JOHN R. SEARLE

EL REDESCUBRIMIENTO DE LA MENTE

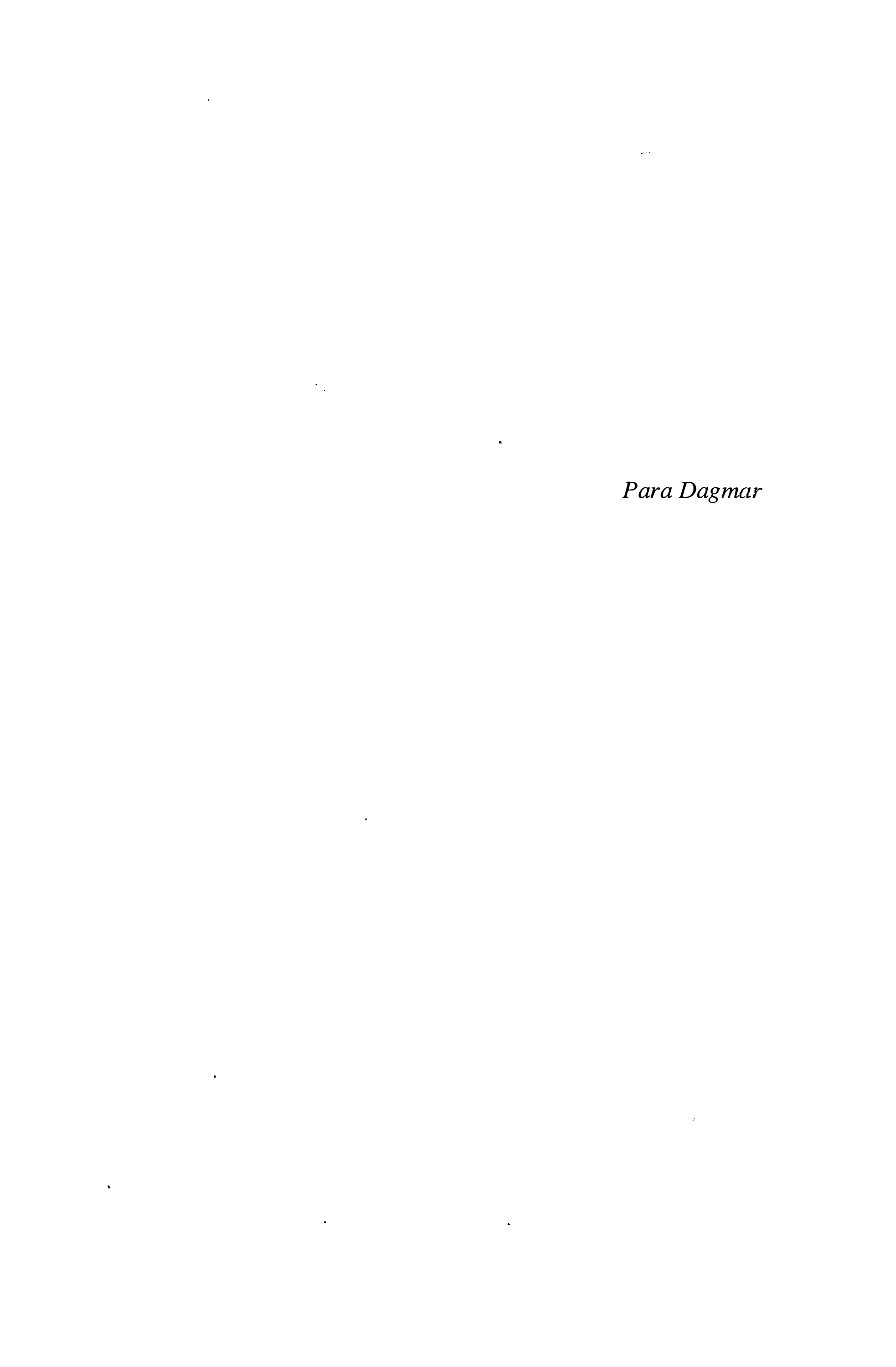
Traducción castellana de
LUIS M. VALDÉS VILLANUEVA

CRÍTICA
GRIJALBO MONDADORI
BARCELONA

Título original:
THE REDISCOVERY OF THE MIND
The MIT Press, Cambridge, Massachusetts

Cubierta: Enric Satué sobre un trabajo artesanal, en pan, de Eduardo Crespo

ISBN: 84-7423-742-4
Depósito legal: B. 36.215-1996
Impreso en España
1996. – NOVAGRÀFIK, S. L., Puigcerdà, 127, 08019 Barcelona



Para Dagmar



AGRADECIMIENTOS

Me he beneficiado, durante un período de varios años, de discusiones y conversaciones con amigos, estudiantes y colegas sobre los problemas considerados en este libro. No supongo que puedo darles las gracias a todos ellos, pero quiero ofrecer expresiones especiales de gratitud a los siguientes: M. E. Aubert, John Batali, Catharine Carlin, Anthony Dardis, Hubert Dreyfus, Hana Filip, Jerry Fodor, Vinod Goel, Stevan Harnad, Jennifer Hudin, Paul Kube, Ernest Lepore, Elisabeth Lloyd, Kirk Ludwig, Thomas Nagel, Randal Parker, Joëlle Proust, Irving Rock, Charles Siewart, Melissa Vaughn y Kayley Vernallis.

Éstos son, sin embargo, sólo algunos de los muchos que tanto me han ayudado. He presentado estas ideas en conferencias que he dado, no solamente en Berkeley, sino también, como profesor visitante, en las universidades de Frankfurt, Venecia, Florencia, Berlín y Rutgers. Mis estudiantes han sido mis mejores y más severos críticos y les estoy agradecido por su incansable escepticismo. Quiero dar las gracias, entre mis benefactores institucionales, al Committee on Research of the Academic Senate y al Office of the Chancellor de la Universidad de California, Berkeley y, especialmente, al Rockefeller Foundation Center en Bellagio, Italia.

Parte del material contenido en este libro ha aparecido en otras partes de una forma preliminar. Específicamente, algunas partes de los capítulos 7 y 10 son un desarrollo de mi artículo «Consciousness, Explanatory Inversion and Cognitive Science» (Behavioral and Brain Sciences, 1990), y el capítulo 9 se basa en mi discurso presidencial a la American Philosophical Association en 1990.

Estoy especialmente agradecido a Ned Block por leer el manuscrito completo —cuando todavía tenía forma de borrador— y hacer mu-

chos comentarios útiles. Y sobre todo quiero dar las gracias a mi esposa, Dagmar Searle, por su constante ayuda y consejo. Como siempre, ella ha ejercido sobre mí la mayor influencia intelectual y ha sido mi más poderosa fuente de ánimo e inspiración. A ella está dedicado este libro.

INTRODUCCIÓN

Este libro tiene varios objetivos, algunos de los cuales no admiten un rápido resumen: sólo emergerán a medida que el lector se vaya sumergiendo en él. Sus objetivos más fácilmente enunciables son estos: quiero criticar y superar las tradiciones dominantes en el estudio de la mente, tanto el «materialismo» como el «dualismo». Puesto que pienso que la conciencia es el fenómeno mental central, quiero comenzar un serio examen de la conciencia en sus propios términos. Quiero poner el último clavo en el ataúd de la teoría de que la mente es un programa de ordenador. Quiero hacer algunas propuestas para reformar nuestro estudio de los fenómenos mentales de una manera que justifique la esperanza de redescubrir la mente.

Hace casi dos décadas comencé a trabajar sobre los problemas de la filosofía de la mente. Necesitaba una explicación de la intencionalidad, para proporcionar tanto un fundamento para mi teoría de los actos de habla, como para completar la teoría. Desde mi punto de vista, la filosofía del lenguaje es una rama de la filosofía de la mente; por consiguiente, ninguna teoría del lenguaje es completa sin una explicación de las relaciones entre mente y lenguaje y de cómo el significado —la intencionalidad derivada de los elementos lingüísticos— está anclado en la intencionalidad intrínseca, biológicamente más básica, de la mente/cerebro.

Cuando leía los autores estándar e intentaba explicar sus puntos de vista a mis estudiantes, me quedaba aterrado al descubrir que, con muy pocas excepciones, esos autores negaban rutinariamente lo que yo pensaba que eran simples y obvias verdades sobre la mente. Era entonces, y es aún muy común, negar, implícita o explícitamente, afirmaciones tales como las siguientes: todos nosotros tenemos estados de conciencia cualitativamente subjetivos, y tenemos estados mentales intrínseca-

mente intencionales tales como creencias y deseos, intenciones y percepciones. Tanto la conciencia como la intencionalidad son procesos biológicos causados por procesos neuronales de nivel más bajo que tienen lugar en el cerebro, y ninguna de las dos cosas es reductible a algo distinto. Además, conciencia e intencionalidad están esencialmente conectadas en el sentido de que la noción de un estado intencional inconsciente solamente la entendemos en términos de su accesibilidad a la conciencia.

Ahora bien, todo esto y más era negado por los puntos de vista dominantes. La principal corriente de la ortodoxia incluye diversas versiones del «materialismo». Los oponentes del materialismo, en un sentido tan rechazable como el anterior, abrazan usualmente alguna doctrina de «dualismo de propiedades», aceptando entonces el aparato cartesiano que, pienso, está desacreditado desde hace bastante tiempo. Lo que he argumentado respecto de ambas posturas (Searle, 1984b) y repito aquí es que uno puede aceptar los hechos obvios de la física —que el mundo consta enteramente de partículas físicas en campos de fuerza— sin negar que entre los rasgos físicos del mundo hay fenómenos biológicos tales como estados de conciencia cualitativamente internos e intencionalidad intrínseca.

Más o menos al mismo tiempo en que comenzó mi interés por los problemas de la mente, nacía la nueva disciplina de la ciencia cognitiva. La ciencia cognitiva prometía una ruptura con la tradición conductista en psicología, puesto que afirmaba entrar en la caja negra de la mente y examinar su funcionamiento interno. Pero, desafortunadamente, muchos científicos cognitivos punteros repitieron simplemente el peor error de los conductistas: insistieron en estudiar solamente fenómenos objetivamente observables, ignorando entonces los rasgos esenciales de la mente. Por consiguiente, cuando abrieron la gran caja negra sólo encontraron en su interior muchas cajas negras pequeñitas.

Así pues, obtuve poca ayuda para mis investigaciones tanto de las corrientes principales de la filosofía de la mente como de la ciencia cognitiva y decidí desarrollar mi propia explicación de la intencionalidad y su relación con el lenguaje (Searle, 1983). Sin embargo, el desarrollar sólo una teoría de la intencionalidad dejaba muchos problemas importantes sin discutir y, peor aún, dejaba sin responder los que parecían ser los errores más importantes y extendidos. Este libro es un intento de llenar, al menos, alguno de esos huecos.

Una de las tareas más difíciles —y más importantes— de la filoso-

fía consiste en clarificar la distinción entre aquellos rasgos del mundo que son *intrínsecos*, en el sentido de que existen independientemente de cualquier observador, y aquellos rasgos que son *relativos al observador*, en el sentido de que sólo existen relativamente a algún observador exterior o usuario. Por ejemplo, el que un objeto tenga una cierta masa es un rasgo intrínseco del objeto. Si todos nosotros muriésemos, el objeto continuaría teniendo masa. Pero que el mismo objeto sea una bañera no es un rasgo intrínseco; existe solamente de manera relativa a usuarios y observadores que le asignan la función de bañera. Tener masa es intrínseco, pero ser una bañera es algo relativo al observador, aun cuando el objeto tenga masa y, a la vez, sea una bañera. Esta es la razón por la que la ciencia natural incluye la masa en su dominio, mientras que no hay ciencia natural de las bañeras.

Uno de los temas que recorre este libro es el intento de clarificar qué predicados de filosofía de la mente nombran rasgos que son intrínsecos y cuáles son relativos al observador. Una postura dominante en filosofía de la mente y en ciencia cognitiva ha sido suponer que la computación es un rasgo intrínseco del mundo y que conciencia e intencionalidad son, de alguna manera, eliminables, bien a favor de algo distinto o porque son relativas al observador, o reductibles a algo más básico, tal como la computación. En este libro argumento que esas suposiciones son exactamente regresivas: conciencia e intencionalidad son intrínsecas e ineliminables y la computación —excepto en los pocos casos en los que la computación se realiza de forma efectiva por una mente consciente— es relativa al observador.

He aquí un mapa conciso para ayudar al lector a encontrar su propio camino en el libro. Los primeros tres capítulos contienen críticas de los puntos de vista dominantes en filosofía de la mente. Son un intento de superar tanto el dualismo como el materialismo; en él se le dedica mayor atención al materialismo. Durante algún tiempo pensé titular a este libro *¿Qué marcha mal en la filosofía de la mente?*, pero al final esta idea emerge como el tema de los primeros tres capítulos y es el título del primero. Los siguientes cinco capítulos, del 4 al 8, son una serie de intentos de dar una caracterización de la conciencia. Una vez que hemos ido más allá tanto del materialismo como del dualismo, ¿cómo colocamos la conciencia en relación con el resto del mundo? (capítulo 4). ¿Cómo damos cuenta de su aparente irreductibilidad de acuerdo con los modelos estándar de la reducción científica? (capítulo 5). Más importante aún: ¿cuáles son los rasgos estructurales de la conciencia?

(capítulo 7). ¿Y cuáles son las relaciones entre conciencia, intencionalidad y las capacidades de «Trasfondo», que nos capacitan para funcionar como seres conscientes en el mundo? (capítulo 8). En el curso de esas discusiones intento superar diversas consignas cartesianas tales como el dualismo de propiedades, el introspeccionismo e incorregibilidad, pero el objetivo principal de estos capítulos no es crítico. Estoy intentando colocar la conciencia dentro de nuestra concepción general del mundo y el resto de nuestra vida mental. El capítulo 9 extiende mis primeras críticas (Searle, 1980a y b) del paradigma dominante en la ciencia cognitiva, y el capítulo final hace algunas sugerencias respecto de cómo podríamos estudiar la mente sin cometer tantos errores obvios.

En este libro tengo más que decir sobre las opiniones de otros autores que en ninguno de mis otros libros; quizás más que en todos ellos juntos. Esto me pone muy nervioso, puesto que siempre es posible que pueda estar interpretándolos de manera tan desastrosa como ellos me interpretan a mí. El capítulo 2 es el que más quebraderos de cabeza me ha dado en este aspecto, y sólo puedo decir que he intentado lo mejor que he podido hacer un resumen imparcial de toda una familia de puntos de vista que me parecen inadmisibles. Por lo que respecta a las referencias: los libros que leí en mi niñez filosófica —libros de Wittgenstein, Austin, Strawson, Ryle, Hare, etc.— contienen pocas referencias —o ninguna— a otros autores. Pienso que, inconscientemente, he llegado a creer que la calidad filosófica varía inversamente con el número de referencias bibliográficas, y que ninguna gran obra de filosofía ha contenido jamás un gran número de notas a pie de página. (Cualesquiera que sean sus otros defectos, *El concepto de lo mental* de Ryle es, en este aspecto, un modelo: no tiene ninguna.) Sin embargo, en este caso no hay posibilidad de escapar a las referencias bibliográficas, y es casi seguro que estoy más en falta por las que he dejado fuera que por las que he tomado en cuenta.

El título es un obvio homenaje al clásico de Bruno Snell, *The Discovery of the Mind*. Ojalá al redescubrir la conciencia —la cosa real, no el sustituto cartesiano ni su doble conductista— redescubramos también la mente.

1. ¿QUÉ MARCHA MAL EN LA FILOSOFÍA DE LA MENTE?

I. LA SOLUCIÓN AL PROBLEMA MENTE-CUERPO Y POR QUÉ MUCHOS PREFIEREN EL PROBLEMA A LA SOLUCIÓN

El famoso problema mente-cuerpo, la fuente de tantas controversias durante los dos últimos milenios, tiene una solución muy simple. Esta solución ha estado al alcance de cualquier persona culta desde que empezaron a realizarse, hace más o menos un siglo, trabajos serios sobre el cerebro y, en un sentido, todos sabemos que es verdadera. Tal solución es la siguiente: los fenómenos mentales están causados por procesos neuropsicológicos del cerebro y son a su vez rasgos del cerebro. Para distinguir este punto de vista de muchos otros que existen en el mercado lo llamaré «naturalismo biológico». Los eventos y procesos mentales son parte de nuestra historia natural biológica en la misma medida en que lo son la digestión, la mitosis, la meiosis o la secreción de enzimas.

El naturalismo biológico plantea por sí mismo miles de cuestiones. ¿Cuál es exactamente el carácter de los procesos neurofisiológicos y cómo producen exactamente los elementos de la neuroanatomía —neuronas, sinapsis, uniones sinápticas, receptores, mitocondrias, células gliales, fluidos transmisores, etc.— fenómenos mentales? ¿Y qué sucede con la enorme variedad de nuestra vida mental —dolores, deseos, cosquilleos, pensamientos, experiencias visuales, creencias, gustos, olores, ansiedad, miedo, amor, odio, depresión y júbilo? ¿Cómo da cuenta la neurofisiología del rango de nuestros fenómenos mentales, tanto conscientes como inconscientes? Tales cuestiones forman el núcleo temático de la neurociencia y en el momento en que escribo esto hay, literalmente, cientos de personas investigando estas cuestio-

nes.[1] Pero no todas las cuestiones son neurobiológicas. Algunas son filosóficas o psicológicas o parte de la ciencia cognitiva entendida de manera general. Algunas de las cuestiones filosóficas son las siguientes: ¿qué es exactamente la conciencia y cómo se relacionan exactamente con el inconsciente los fenómenos mentales conscientes? ¿Cuáles son los rasgos especiales de lo «mental», rasgos tales como conciencia, intencionalidad, subjetividad, causación mental? ¿Cómo funcionan exactamente? ¿Cuáles son las relaciones causales entre fenómenos «mentales» y «físicos»? ¿Y podemos caracterizar esas relaciones causales de una manera que evite el epifenomenalismo?

Intentaré decir algo sobre algunas de estas cuestiones más adelante, pero en este punto quiero resaltar un hecho destacable. He dicho que la solución al problema mente-cuerpo debería de ser obvia para una persona culta, pero, en la actualidad, muchos de los expertos, quizás la mayoría, en filosofía y ciencia cognitiva afirman que no la encuentran obvia en absoluto. De hecho, ni siquiera piensan que la solución que he propuesto sea verdadera. Si se revisa el campo de la filosofía de la mente durante las últimas décadas, nos encontramos con que dicho campo está ocupado por una pequeña minoría que insiste en la realidad e irreductibilidad de la conciencia y la intencionalidad y que tiende a pensarse como dualista de propiedades, y un grupo dominante mucho más amplio que se piensan a sí mismos como materialistas de uno u otro tipo. Los dualistas de propiedades piensan que el problema mente-cuerpo es aterradoramente difícil, quizás totalmente irresoluble.[2] Los materialistas están de acuerdo en que si la intencionalidad y la conciencia existen realmente y son irreductibles a los fenómenos físicos, entonces tendríamos realmente un difícil problema mente-cuerpo, pero esperan «naturalizar» la intencionalidad y quizás también la conciencia. Por «naturalizar» los fenómenos mentales entienden el reducirlos a fenómenos físicos. Piensan que aceptar la realidad e irreductibilidad de la conciencia y de otros fenómenos mentales nos compromete con alguna forma de cartesianismo, y no ven cómo tal punto de vista puede hacerse consistente con nuestra representación científica global del mundo.

1. O, al menos, están investigando los preliminares de tales cuestiones. Resulta sorprendente la proporción tan pequeña de la neurociencia que está dedicada a investigar, por ejemplo, la neurofisiología de la conciencia.

2. El proponente mejor conocido de este punto de vista es Thomas Nagel (1986), pero véase también Colin McGinn (1991).

Creo que ambas partes están profundamente equivocadas. Ambas aceptan cierto vocabulario y con él un conjunto de supuestos. He intentado mostrar que el vocabulario es obsoleto y que los supuestos son falsos. Es esencial mostrar que tanto el dualismo como el monismo son falsos puesto que, generalmente, se supone que ambos agotan el campo y no dejan otra opción. Muchas de mis discusiones estarán dirigidas a las diversas formas de materialismo puesto que se trata del punto de vista dominante. El dualismo de cualquier forma se considera hoy día de manera general como algo fuera de toda consideración puesto que se supone que es inconsistente con la visión científica del mundo.

Así pues, la cuestión que quiero plantear en este capítulo y en el próximo es la siguiente: ¿qué pasa en nuestra historia intelectual y en nuestro entorno que hace difícil ver estas puntualizaciones más bien simples que he hecho sobre el «problema mente-cuerpo»? ¿Qué ha hecho que el «materialismo» aparezca como el único enfoque racional en filosofía de la mente? Este capítulo y el siguiente tratan sobre la situación actual en filosofía de la mente, y éste podría haberse titulado «¿Qué marcha mal en la tradición materialista en filosofía de la mente?».

Vista desde la perspectiva de los últimos cincuenta años, la filosofía de la mente, así como la ciencia cognitiva y ciertas ramas de la psicología, presenta un espectáculo muy curioso. El rasgo más sorprendente es la enorme cantidad de filosofía de la mente dominante en los últimos cincuenta años que parece obviamente falsa. Creo que no hay ninguna otra área de la filosofía analítica contemporánea donde se haya dicho tanto que sea tan implausible. En la filosofía del lenguaje, por ejemplo, no es común en absoluto negar la existencia de oraciones y actos de habla; pero en la filosofía de la mente hechos obvios sobre lo mental, tales como que todos tenemos realmente estados mentales subjetivos conscientes y que éstos no son eliminables en favor de algo distinto, se niegan de manera rutinaria por muchos, quizás por la mayoría, de los pensadores más avanzados sobre el tema.

¿Cómo es que tantos filósofos y científicos cognitivos pueden decir tantas cosas que, a mí al menos, me parecen obviamente falsas? Los puntos de vista extremos en filosofía no son casi nunca carentes de inteligencia; hay generalmente razones muy profundas y poderosas para mantenerlos. Creo que uno de los supuestos no enunciados que subyace en la actual hornada de puntos de vista es que éstos repre-

sentan las únicas alternativas científicamente aceptables al anticientifismo que conllevaba el dualismo tradicional, la creencia en la inmortalidad del alma, el espiritualismo y cosas por el estilo. La aceptación de los puntos de vista actuales está motivada no tanto por una convicción independiente de que son verdaderos como por un terror a lo que, aparentemente, son las únicas alternativas. Esto es: la elección ante la que tácitamente se nos pone es entre un enfoque «científico», tal como el que viene representado por una u otra de las versiones actuales del «materialismo», y un enfoque «anticientífico», tal como el que viene representado por el cartesianismo o alguna otra concepción religiosa tradicional de la mente. Otro hecho extraño, estrechamente relacionado con el primero, es que muchos de los autores estándar están profundamente comprometidos con las categorías y el vocabulario tradicionales. Piensan realmente que hay un significado más o menos claro que va ligado al vocabulario arcaico de «dualismo» «monismo», «materialismo», «fisicalismo», etc., y que los problemas tienen que plantearse y resolverse en estos términos. Usan esas palabras sin embarazo ni ironía alguna. Una de las principales aspiraciones que tengo en este libro es mostrar que ambos supuestos son erróneos. Entendidos apropiadamente, muchos de los puntos de vista actualmente en boga son inconsistentes con lo que sabemos sobre el mundo, tanto a partir de nuestras propias experiencias como a partir de las ciencias especiales. Para enunciar lo que todos sabemos que es verdad, tendremos que desafiar los supuestos que subyacen en el vocabulario tradicional.

Antes de identificar algunos de estos increíbles puntos de vista, quiero hacer alguna observación sobre el estilo de presentarlos. Los autores que van a decir algo que suena estúpido muy a menudo se topan con ello y lo dicen. Usualmente se emplea un conjunto de dispositivos retóricos o estilísticos para evitar decirlo en palabras de una sílaba. El más obvio de esos dispositivos consiste en marear la perdiz con una gran cantidad de prosa evasiva. Pienso que resulta obvio en los escritos de diversos autores que, por ejemplo, piensan que no tenemos realmente estados mentales, tales como creencias, deseos, temores, etc. Pero es difícil encontrar pasajes donde digan esto claramente. A menudo intentan mantener el vocabulario de sentido común, mientras niegan que esté efectivamente por algo en el mundo real. Otro dispositivo retórico para disfrazar lo implausible es darle un nombre al punto de vista del sentido común y, a continuación, negarlo por el nombre y no

por el contenido. Efectivamente, es muy difícil, incluso en los tiempos presentes, llegar y decir: «Ningún ser humano ha sido consciente jamás». Más bien, el filósofo sofisticado da un nombre al punto de vista de que la gente es consciente algunas veces, por ejemplo, «la intuición cartesiana». Una vez más, es difícil decir que nadie en la historia del mundo bebió jamás porque estaba sediento o comió porque estaba hambriento; pero es fácil desafiar algo si se lo puede rotular de antemano como «psicología popular». Y a los únicos efectos de dar un nombre a esta maniobra, la llamaré la maniobra de «dar-le-un-nombre». Llamaré a otra de las maniobras, la más favorita de todas, la maniobra de la «edad-heroica-de-la-ciencia». Cuando un autor se encuentra en una dificultad profunda, él o ella intenta hacer una analogía entre su propia afirmación y algún gran descubrimiento científico del pasado. ¿Parece estúpido este punto de vista? Bien, los grandes genios científicos del pasado parecían estúpidos a sus contemporáneos, unos ignorantes, dogmáticos y llenos de prejuicios. Galileo es la analogía histórica favorita. Retóricamente hablando, la idea es hacer que usted, el lector escéptico, sienta que si no cree el punto de vista que se está avanzando, está representando el papel del cardenal Belarmino mientras que el autor representa el de Galileo.[3] Otras maniobras favoritas son el *flogisto* y los *espíritus vitales*, y la idea es, de nuevo, amedrentar al lector con la suposición de que si él o ella dudan de que, por ejemplo, los ordenadores piensan efectivamente, esto sólo puede deberse a que el lector cree en algo tan poco científico como el flogisto o los espíritus vitales.

II. SEIS TEORÍAS INVEROSÍMILES DE LA MENTE

No voy a intentar proporcionar un catálogo completo de todas las visiones materialistas, tan en boga y, a la vez, tan implausibles, que se nos ofrecen en la filosofía y en la ciencia cognitiva contemporáneas, pero haré una relación de sólo una media docena de ellas para tener una percepción directa del asunto. Lo que esas visiones comparten es una hostilidad hacia la existencia y el carácter mental de nuestra vida mental ordinaria. De una manera u otra, todas ellas intentan degradar los fenómenos mentales ordinarios tales como creencias, deseos e intenciones y arrojar

3. Véase, por ejemplo, P. S. Churchland (1987).

dudas sobre la existencia de rasgos generales de los fenómenos mentales tales como la conciencia y la subjetividad.[4]

En primer lugar, quizás la versión más extrema de esos puntos de vista es la idea de que los estados mentales, como tales, no existen en absoluto. Este punto de vista es mantenido por aquellos que se llaman a sí mismos «materialistas eliminativos». La idea es que, contrariamente a una creencia muy extendida, no hay en realidad cosas tales como creencias, deseos, esperanzas, temores, etc. Las primeras versiones de este punto de vista fueron avanzadas por Feyerabend (1963) y Rorty (1965).

Un segundo punto de vista, usado a menudo para apoyar el materialismo eliminativo, es la afirmación de que la psicología popular es —con toda probabilidad— simple y enteramente falsa. Este punto de vista ha sido avanzado por P. M. Churchland (1981) y Stich (1983). La psicología popular incluye afirmaciones tales como que las personas beben algunas veces porque están sedientas y comen porque tienen hambre; que tienen deseos y creencias, que algunas de esas creencias son verdaderas o, cuando menos, falsas; que algunas creencias están mejor fundadas que otras; que las personas hacen algunas veces cosas porque quieren hacerlas; que algunas veces tienen dolores; y que esos dolores son muy a menudo desagradables. Y así sucesivamente, de manera más o menos indefinida. La conexión entre la psicología popular y el materialismo eliminativo es la siguiente: se supone que la psicología popular es una teoría empírica y que las «entidades» que postula —dolores, cosquilleos, picores, y cosas por el estilo— son entidades teóricas, ontológicamente hablando, por los cuatro costados, del mismo modo que los quarks o los muones. Si la teoría cae, las entidades teóricas van con ella: demostrar la falsedad de la psicología popular sería eliminar cualquier justificación para aceptar la existencia de entidades psicológicas populares. Espero sinceramente no ser injusto al caracterizar como implausibles esos puntos de vista, pero tengo que confesar que este es el modo como me parecen las cosas. Continuemos con la lista.

Un tercer punto de vista de este mismo tipo mantiene que no hay

4. Limitaré mi discusión a los filósofos analíticos, pero, aparentemente, el mismo tipo de implausibilidad afecta a la llamada filosofía continental. De acuerdo con Dreyfus (1991), Heidegger y sus seguidores dudan también de la importancia de la conciencia y la intencionalidad.

nada que sea específicamente *mental* en los llamados estados mentales. Los estados mentales consisten enteramente en sus relaciones causales entre sí y con los *inputs* y *outputs* del sistema del que son parte. Esas relaciones causales podrían duplicarse en cualquier sistema que tuviese las propiedades causales correctas. Así pues, un sistema hecho de piedras o latas de cerveza, si tuviese las relaciones causales correctas, tendría que tener las mismas creencias, deseos, etc., que tenemos nosotros, puesto que esto es todo aquello en lo que consiste tener creencias y deseos. La versión más influyente de este punto de vista se denomina «funcionalismo», y se acepta tan ampliamente que constituye una de las ortodoxias contemporáneas.

Un cuarto punto de vista implausible, de hecho el más famoso y más ampliamente mantenido del actual catálogo, es el de que un ordenador podría tener —de hecho tiene que tener— pensamientos, sentimientos, y comprensión en virtud solamente de la implementación de un programa apropiado de ordenador con los *inputs* y *outputs* apropiados. En otro lugar, he bautizado este punto de vista como «inteligencia artificial fuerte», pero también se le ha denominado «funcionalismo de ordenador».

Una quinta forma de visión increíble se halla en la afirmación de que no deberíamos pensar en nuestro vocabulario mental de «creencia», «deseo», «temor» y «esperanza», etc., como algo que representa fenómenos intrínsecamente mentales, sino más bien como una manera de hablar. Se trataría solamente de un vocabulario útil para explicar y predecir la conducta, pero no debería tomarse literalmente como si hiciese referencia a fenómenos psicológicos subjetivos, intrínsecos, reales. Aquellos que se adhieren a este punto de vista piensan que el uso del vocabulario del sentido común es asunto de adoptar una «postura intencional» hacia un sistema.[5]

En sexto lugar, otro punto de vista extremo es que la conciencia, tal como nosotros pensamos en ella —como fenómenos cualitativos de sentir o darse cuenta de manera interna, privada y subjetiva— no existe en absoluto. Este punto de vista se avanza explícitamente muy pocas veces.[6] Muy poca gente quiere decir lisa y llanamente que la conciencia no existe. Pero, recientemente, se ha convertido en algo común entre ciertos autores redefinir la noción de conciencia de modo que ya no

5. El exponente mejor conocido de este punto de vista es Daniel Dennett (1987).
6. Véase, para un enunciado explícito del mismo, Georges Rey (1983).

se refiera a estados conscientes efectivos, esto es: a estados mentales de primera persona, cualitativos, subjetivos, internos, sino más bien a fenómenos de tercera persona públicamente observables. Tales autores aparentan pensar que la conciencia existe, pero, de hecho, terminan negando su existencia.[7]

Algunas veces, los errores en filosofía de la mente producen errores en filosofía del lenguaje. Una tesis, a mi juicio increíble, de filosofía del lenguaje, que es del mismo estilo que los ejemplos que hemos estado considerando, es el punto de vista de que, por lo que respecta a los significados, no hay hecho objetivo alguno además de los modelos de conducta verbal. De acuerdo con esta posición, mantenida notoriamente por Quine (1960), no hay ningún hecho objetivo respecto de si cuando tú o yo decimos «conejo» queremos decir conejo, parte no separada de conejo, o estado en la historia de la vida de un conejo.[8]

Ahora bien, ¿qué ha de hacerse ante todo esto? No me resulta suficiente decir que todo ello parece implausible; pienso más bien que un filósofo con paciencia y tiempo suficientes debería sentarse y hacer una refutación, línea por línea, de toda esa tradición. He intentado hacer esto con una tesis específica de esta tradición: la afirmación de que los ordenadores tienen pensamientos, sentimientos y comprensión en virtud solamente de instanciar un programa de ordenador (el programa de ordenador «correcto» con los *inputs* y los *outputs* «correctos») (Searle, 1980a). Este punto de vista, la inteligencia artificial fuerte, ofrece un blanco atractivo puesto que está razonablemente claro que existe una refutación simple y decisiva, y la refutación puede extenderse a otras versiones del funcionalismo. He intentado también refutar la tesis de Quine de la indeterminación (Searle, 1987), que creo que también se presta a un asalto frontal. Con alguno de los puntos de vista la situación está, sin embargo, mucho más embrollada. ¿Cómo, por ejemplo, procedería uno a refutar la posición de que la conciencia no existe? ¿Debería pellizcar a los que la mantienen para recordarles que son cons-

7. Creo que esto lo hacen, de diferentes maneras, Armstrong (1968, 1980), y Dennett (1991).

8. Otra forma increíble, pero desde una motivación filosófica diferente, es la afirmación de que cada uno de nosotros tiene, desde su nacimiento, todos los conceptos expresables en cualesquiera palabras de cualquier lenguaje humano posible, de modo que, por ejemplo, los hombres de Cro-Magnon tenían los conceptos expresables por la palabra «carburador» o por la expresión «oscilógrafo de rayos catódicos». Esta posición es mantenida notoriamente por Fodor (1975).

cientes? ¿Debería pellizcarme a mí mismo e informar de los resultados en el *Journal of Philosophy*?

Para desarrollar un argumento, en el sentido tradicional, es necesario que haya alguna base común. A menos que los participantes estén de acuerdo en las premisas, no tiene objeto intentar derivar una conclusión. Pero si alguien niega, desde el principio, la existencia de la conciencia, es difícil saber cuál sería la base común para el estudio de la mente. De acuerdo con mi punto de vista, si nuestra teoría da como resultado la posición de que la conciencia no existe, entonces lo que que se ha producido es, simplemente, una reducción al absurdo de la teoría y la situación es similar en muchos otros puntos de vista de la filosofía contemporánea de la mente.

Los diversos años de debate de esos problemas, tanto en foros públicos como en publicaciones, me han convencido de que, muy a menudo, los problemas fundamentales del debate no salen a la superficie. Si se debate con gente acerca de, por ejemplo, la inteligencia artificial fuerte o la indeterminación de la traducción, la pura y simple implausibilidad de tales teorías se disfraza con el carácter aparentemente técnico de los argumentos esgrimidos una y otra vez. Peor aún, es difícil sacar a la luz las suposiciones que llevan a esas teorías. Cuando, por ejemplo, alguien se siente a gusto con la idea de que un ordenador podría tener, de repente y de modo milagroso, estados mentales solamente en virtud de ejecutar cierta suerte de programa, las suposiciones subyacentes que hacen que esta posición parezca plausible raramente se enuncian de modo explícito. Así pues, en esta exposición quiero intentar un enfoque diferente del de la refutación directa. No voy a presentar una o más de una «refutaciones del funcionalismo»; más bien, lo que quiero es dar comienzo a la tarea de exponer y, mediante ello, socavar los cimientos sobre los que descansa la totalidad de esta tradición. Si a usted le tienta el funcionalismo, creo que lo que usted necesita no es una refutación, lo que usted necesita es ayuda.

La tradición materialista es sólida, compleja, ubicua y, con todo, evasiva. Sus diversos elementos —su actitud hacia la conciencia, su concepción de la verificación científica, su metafísica y su teoría del conocimiento— se apoyan mutuamente, de modo que cuando se desafía una parte, los defensores pueden fácilmente echar mano de otra parte cuya certeza se da por sentada. Estoy hablando aquí de mi experiencia personal. Cuando se ofrece una refutación de la Inteligencia Artificial (IA) fuerte o de la tesis de la indeterminación o del funciona-

lismo, los defensores no perciben que sea necesario intentar hacer frente a tus argumentos efectivos, puesto que saben de antemano que tú tienes que estar equivocado. Saben que la tradición materialista —que ellos llaman a menudo, y erróneamente, «ciencia»— está de su parte. Y la tradición no es sólo parte de la filosofía académica. Si uno escucha conferencias sobre ciencia cognitiva o lee artículos de divulgación sobre inteligencia artificial, se encontrará con la misma tradición. Esto es algo demasiado extenso para resumirlo en un párrafo o ni siquiera en un capítulo, pero creo que si continúo dejando que se despliegue por sí mismo, el lector no tendrá dificultad en reconocerlo.

Antes de comenzar el asalto a los cimientos, necesito especificar ciertos elementos de la estructura de manera un poco más precisa y decir algo sobre su historia.

III. Los fundamentos del materialismo moderno

Por «la tradición», entiendo en gran parte el conjunto de puntos de vista y presuposiciones metodológicas que se centran en torno a los siguientes (a menudo no enunciados) supuestos y tesis:

1. Por lo que respecta al estudio científico de la mente, la conciencia y sus rasgos especiales son, más bien, de menor importancia. Es completamente posible, de hecho es deseable, dar una explicación del lenguaje, de la cognición y de los estados mentales en general sin tomar en cuenta la conciencia y la subjetividad.[9]

2. La ciencia es objetiva. Es objetiva no sólo en el sentido de que intenta alcanzar conclusiones que son independientes de sesgos personales y puntos de vista, sino, y esto es más importante, que se interesa por una realidad que es objetiva. La ciencia es objetiva porque la realidad misma es objetiva.

3. Puesto que la realidad es objetiva, el mejor método para estudiar la mente es adoptar el punto de vista objetivo o de tercera persona. La objetividad de la ciencia exige que los fenómenos estudiados sean completamente objetivos y en el caso de la ciencia cognitiva esto significa que tiene que estudiar *conducta* objetivamente observable. Por lo

9. Howard Gardner, en su resumen general de la ciencia cognitiva (1985), no incluye un solo capítulo —de hecho, ni siquiera una simple entrada en el índice— sobre la conciencia. Claramente, la nueva ciencia de la mente puede arreglárselas sin la conciencia.

que respecta a la ciencia cognitiva madura, el estudio de la mente y el estudio de la conducta inteligente (incluyendo los fundamentos causales de la conducta) son totalmente el mismo estudio.

4. Desde el punto de vista objetivo de la tercera persona, la única respuesta a la cuestión epistemológica «¿Cómo conoceríamos los fenómenos mentales de otro sistema?» es la siguiente: los conocemos observando su *conducta.* Esta es la única solución al «problema de las otras mentes».

La epistemología juega un papel especial en la ciencia cognitiva puesto que una ciencia objetiva de la cognición debe ser capaz de distinguir cosas tales como *cognición, conducta inteligente, procesamiento de la información,* etc., de otros fenómenos naturales. Una cuestión básica, quizás la cuestión básica, en el estudio de la mente es la siguiente cuestión epistemológica: ¿Cómo sabríamos si algún otro «sistema» tiene o no tales-y-cuales propiedades mentales? Y la única respuesta científica es: mediante su conducta.

5. La conducta inteligente y las relaciones causales con la conducta inteligente son, de algún modo, la esencia de lo mental. La adhesión al punto de vista de que hay una conexión esencial entre mente y conducta tiene un rango que va desde la versión extrema del conductismo, que dice que no hay nada en lo que consista tener estados mentales excepto el tener disposiciones para la conducta, pasando por los intentos funcionalistas de definir las nociones mentales en términos de relaciones causales externas e internas, hasta la problemática afirmación de Wittgenstein (1953, parágrafo 580) de que «Un "proceso interno" necesita criterios externos».[10]

6. Todo hecho del universo es, en principio, cognoscible y entendible por investigadores humanos. Puesto que la realidad es física, y puesto que la ciencia se interesa por la investigación de la realidad física, y puesto que no hay límites a lo que podemos conocer de la realidad física, se sigue que todos los hechos del universo son cognoscibles y entendibles por nosotros.

7. Las únicas cosas que existen son, en último término, físicas, *tal como la física se concibe tradicionalmente*, esto es: como opuesto a lo mental. Esto significa que en las oposiciones tradicionales —dualismo *versus* monismo, mentalismo *versus* materialismo— el término del

10. De acuerdo con mi punto de vista, un proceso interno no «necesita» nada. ¿Por qué habría de necesitarlo?

lado derecho nombra el punto de vista correcto; el de la parte izquierda, el falso.

Debería estar claro ya que estos puntos de vista van unidos; puesto que la realidad es *objetiva* (punto 2), tiene que ser, en última instancia, *física* (punto 7). Y la ontología objetivista de los puntos 2 y 7 lleva de manera natural a la metodología objetivista de los puntos 3 y 4. Pero si la mente existe realmente y tiene una ontología objetiva, entonces parece que su ontología tiene que ser en algún sentido conductista y causal (punto 5). Esto, sin embargo, fuerza a la epistemología a situarse en primera fila (punto 4), puesto que ahora se convierte en algo crucialmente importante el ser capaces de distinguir la conducta de los sistemas que carecen de estados mentales de aquellos que realmente tienen estados mentales. Del hecho de que la realidad es, en última instancia, física (punto 7), y del hecho de que es completamente objetiva (punto 2), es natural suponer que, en realidad, todo es cognoscible por nosotros (punto 6). Finalmente, una cosa es obvia: no hay lugar en este cuadro general —o, como máximo, hay poquísimo lugar— para la conciencia (punto 1).

A lo largo de este libro espero mostrar que cada uno de estos puntos es, en el mejor de los casos, falso, y que el cuadro total que presentan no sólo es profundamente acientífico, es incoherente.

IV. Orígenes históricos de los fundamentos

¿Cómo hemos llegado, históricamente hablando, a esta situación? ¿Cómo hemos llegado a una situación en la que se dicen cosas que son incompatibles con hechos obvios de sus experiencias?

Lo que uno quiere saber es lo siguiente: ¿qué ha pasado en la historia de las discusiones contemporáneas sobre filosofía de la mente, psicología, ciencia cognitiva e inteligencia artificial que hace que tales puntos de vista sean concebibles, que hace que parezcan perfectamente respetables o aceptables? En cualquier tiempo dado de la historia intelectual, todos nosotros estamos trabajando dentro de ciertas tradiciones que hacen que ciertas preguntas parezcan ser las preguntas que han de plantearse y ciertas respuestas parezcan las únicas respuestas posibles. En la filosofía de la mente contemporánea, la tradición histórica nos ciega para los hechos obvios de nuestras experiencias y nos da una metodología y un vocabulario que hace que

hipótesis que son obviamente falsas parezcan aceptables. La tradición ha ido creciendo desde sus primeros y crudos comienzos conductistas hace ya más de medio siglo, pasando por las teorías de la identidad «tipo-tipo» e «instancia-instancia» hasta los actuales y sofisticados modelos de cognición computacionales. Ahora bien, ¿qué sucede con una tradición que hace esto tan poderoso de una manera tan contraintuitiva? Desearía haber entendido estos asuntos de manera suficiente para dar un análisis histórico completo, pero temo tener sólo un puñado de conjeturas que hacer sobre la naturaleza de los síntomas. Me parece que hay, al menos, cuatro factores implicados en este asunto.

En primer lugar, tenemos terror a caer en el dualismo cartesiano. La bancarrota de la tradición cartesiana, y el absurdo de suponer que hay dos géneros de substancias o propiedades en el mundo, «mentales» y «físicas», nos intimida de tal manera y tiene una historia tan sórdida que somos muy renuentes a aceptar cualquier cosa que pudiese tener un regusto cartesiano. Somos renuentes a aceptar cualquiera de los hechos de sentido común que suenan a «cartesianismo», porque parece que, si aceptamos los hechos, tendremos que aceptar la totalidad de la metafísica cartesiana. Cualquier género de mentalismo que reconozca los hechos obvios de nuestra existencia se considera automáticamente como sospechoso. En el extremo del todo, algunos filósofos son renuentes a admitir la existencia de la conciencia porque no logran ver que el estado *mental* de conciencia es sólo un rasgo biológico ordinario, esto es, *físico*, del cerebro. Quizás estén ayudados, de manera totalmente exasperante, por aquellos filósofos que reconocen alegremente la existencia de la conciencia y, al hacer esto, suponen que tiene que estar aseverando la existencia de algo no físico.

El punto de vista de que la conciencia, los estados mentales, etc., existen, en el sentido más ingenuo y obvio, y juegan un papel causal real en nuestra conducta, no tiene nada especial que ver con el dualismo cartesiano. Despues de todo, uno no tiene que leer las *Meditaciones* para ser consciente de que uno es consciente, o de que los propios deseos, como fenómenos mentales, conscientes o inconscientes, son fenómenos causales reales. Pero cuando uno recuerda a los filósofos esas «intuiciones cartesianas», es acusado inmediatamente de cartesianismo. Yo mismo, hablando personalmente, he sido acusado de mantener alguna loca doctrina de «dualismo de propiedades» y «acceso privilegiado», o de creer en la «introspección» o en el «neovitalismo» o in-

cluso en el «misticismo», aun cuando jamás he apoyado, implícita o explícitamente, ninguno de esos puntos de vista. ¿Por qué? En parte, sin duda, se debe simplemente a una falta de cuidado intelectual (o quizás incluso a algo peor) por parte de los comentadores, pero hay también algo más profundo que está involucrado aquí. Encuentran difícil ver que uno podría aceptar los hechos obvios sobre los estados mentales sin aceptar el aparato cartesiano que, tradicionalmente, ha acompañado el conocimiento de esos hechos. Piensan que las únicas elecciones reales disponibles son alguna forma de materialismo y alguna forma de dualismo. Una de las aspiraciones que tengo al escribir este libro es mostrar que esta concepción es errónea, que uno puede proporcionar una explicación coherente de los hechos sobre la mente sin apoyar nada del desacreditado aparato cartesiano.

En segundo lugar, junto con la tradición cartesiana, hemos heredado un vocabulario, y con el vocabulario un cierto conjunto de categorías dentro de las que estamos históricamente condicionados a pensar sobre esos problemas. El vocabulario no es inocente, puesto que en el vocabulario están implícitas un número sorprendente de afirmaciones teóricas que son, casi con certeza, falsas. El vocabulario incluye una serie de oposiciones aparentes «físico» *versus* «mental», «cuerpo» *versus* «mente», «materialismo» *versus* «mentalismo», «materia» *versus* «espíritu». En estas oposiciones está implícita la tesis de que el mismo fenómeno bajo los mismos aspectos no puede satisfacer literalmente los dos términos. Algunas veces la semántica, e incluso la morfología, parecen hacer explícitas estas observaciones, como sucede en la aparente oposición entre «materialismo» e «inmaterialismo». Así pues, se supone que creemos que si algo es mental, no puede ser físico; que si es un asunto del espíritu, no puede serlo de la materia; si es inmaterial, no puede ser material. Pero estos puntos de vista me parecen obviamente falsos, dado todo lo que sabemos sobre la neurobiología. El cerebro causa ciertos fenómenos «mentales», tales como los estados mentales conscientes, y esos estados conscientes son, simplemente, rasgos de nivel superior del cerebro. La conciencia es una propiedad emergente, o de nivel superior, del cerebro en el sentido lisa y llanamente inocuo de «nivel superior» y «emergente» en el que la solidez es una propiedad emergente de nivel superior de las moléculas de H_2O cuando están en una estructura de enrejado (hielo), y la liquidez es, de manera similar, una propiedad emergente de nivel superior de las moléculas de H_2O cuando están, para decirlo de manera aproximada, rodando unas con

otras (agua). La conciencia es una propiedad mental y, por lo tanto, física, del cerebro en el sentido en que la liquidez es una propiedad de sistemas de moléculas. Si hay una tesis que quisiera mantener en esta discusión es, simplemente, esta: el hecho de que un rasgo es mental no implica que no sea físico; el hecho de que un rasgo e físico no implica que no sea mental. Revisando en este momento a Descartes podríamos decir, no sólo «Pienso, luego existo» y «Soy un ser que piensa», sino también *Soy un ser pensante, luego soy un ser físico.*

Pero obsérvese cómo el vocabulario hace difícil, si no imposible, decir lo que quiero decir usando la terminología tradicional. Cuando digo que la conciencia es un rasgo de nivel superior del cerebro, la tentación es oír esto como queriendo decir físico-como-opuesto-a-mental, como queriendo decir que la conciencia debería describirse *sólo* en términos conductistas objetivos o neurofisiológicos. Pero lo que realmente yo quiero decir es que la conciencia *qua* conciencia, *qua* mental, *qua* subjetiva, *qua* cualitativa, es *física*, y es física *porque* es mental. Todo esto muestra, creo, la inadecuación del vocabulario tradicional.

Junto con las oposiciones aparentes están los nombres que, aparentemente, agotan las posiciones posibles que pueden ocuparse: están el monismo *versus* el dualismo, el materialismo y el fisicalismo *versus* el mentalismo y el idealismo. La buena disposición a mantenerse aferrado a las categorías tradicionales produce alguna terminología extraña, tal como el «dualismo de propiedades», el «monismo anómalo», la «identidad como instancia», etc. Mis propios puntos de vista no encajan en ninguna de las etiquetas tradicionales, pero para muchos filósofos, la idea de que se podría mantener un punto de vista que no encaje con esas categorías parece incomprensible.[11] Lo peor de todo es quizás que hay diversos nombres y verbos que parece como si tuvieran un significado claro y representasen efectivamente objetos y actividades bien definidas —«mente», «yo» e «introspección» son ejemplos obvios. El vocabulario de la ciencia cognitiva contemporánea no es mejor. Tendemos a suponer de manera acrítica que expresiones como «cognición», «inteligencia» y «procesamiento de la información» tienen definicio-

11. De manera bastante extraña, mis puntos de vista han sido caracterizados con toda confianza por algunos comentaristas como «materialistas»; por algunos otros, con igual confianza, como «dualistas». Así, por ejemplo, U. T. Place escribe que Searle «presenta la posición materialista» (1988, p. 208), mientras que Stephen P. Stich escribe: «Searle es un dualista de propiedades» (1987, p. 133).

nes claras y representan efectivamente algunos géneros naturales. Creo que tales suposiciones son erróneas. Merece la pena subrayar este punto: «inteligencia», «conducta inteligente», «cognición» y «procesamiento de la información», por ejemplo, no son nociones definidas de modo preciso. Incluso, más sorprendentemente, gran cantidad de nociones que suenan a técnicas están muy pobremente definidas —nociones tales como «ordenador», «computación», «programa» y «símbolo», por ejemplo. En las ciencias de la computación no importa para muchos propósitos que estas nociones estén mal definidas (lo mismo que no es importante tampoco para los fabricantes de muebles que no tengan una definición filosóficamente precisa de «silla» o de «mesa»); pero cuando los científicos cognitivos dicen cosas tales como que los cerebros son ordenadores, las mentes son programas, etc., entonces la definición de esas nociones se convierte en crucial.

En tercer lugar, hay una persistente tendencia objetivadora en la filosofía contemporánea, en la ciencia y en la vida intelectual en general. Tenemos la convicción de que, si algo es real, tiene que ser igualmente accesible a todos los observadores competentes. Desde el siglo XVII, las personas cultas de Occidente han venido aceptando una presuposición metafísica absolutamente básica: *la realidad es objetiva.* Esta suposición ha mostrado que nos resulta útil en muchos aspectos, pero es obviamente falsa como revela un solo momento de reflexión sobre los propios estados subjetivos. Y este supuesto ha llevado, quizás inevitablemente, al punto de vista de que el único modo «científico» de estudiar la mente es considerarla como un conjunto de fenómenos objetivos. Una vez que adoptamos el supuesto de que cualquier cosa que es objetiva debe de ser igualmente accesible a cualquier observador, las cuestiones pasan inmediatamente de la subjetividad de los estados mentales hacia la objetividad de la conducta externa. Y esto tiene la consecuencia de que en vez de plantear las preguntas: «¿Qué es tener una creencia?», «¿Qué es tener un deseo?», «¿Qué es estar en ciertas clases de estados conscientes?», planteamos la cuestión de tercera persona: «¿Bajo qué condiciones *atribuiríamos* desde fuera creencias, deseos, etc., a algún *otro* sistema?». Esto nos parece perfectamente natural puesto que, desde luego, muchas de las cuestiones que necesitamos responder sobre los fenómenos mentales conciernen a otras personas y no sólo a nosotros mismos.

Pero el carácter de tercera persona de la epistemología no debería

cegarnos para el hecho de que la ontología efectiva de los estados mentales es una ontología de primera persona. El modo en que se aplica en la práctica el punto de vista de la tercera persona hace difícil ver la diferencia entre algo que tiene en realidad una mente, tal como un ser humano, y algo que se comporta *como si* tuviera una mente, tal como un ordenador. Y una vez que se pierde la distinción entre que un sistema tenga realmente estados mentales y que actúe meramente como si tuviera estados mentales, entonces se pierde de vista un rasgo esencial de lo mental, a saber: que su ontología es esencialmente una ontología de primera persona. Creencias, deseos, etc., son siempre las creencias y deseos *de alguien*, y son siempre potencialmente conscientes, incluso en los casos en los que son efectivamente inconscientes.

Presento un argumento a favor de este último punto en el capítulo 7. Ahora estoy intentando diagnosticar un modelo de investigación históricamente condicionado que hace que el punto de vista de la tercera persona parezca el único punto de vista aceptable a partir del cual examinar la mente. Sería una tarea de un historiador intelectual el responder a preguntas tales como ¿cuándo la pregunta sobre bajo-qué-condiciones-atribuimos-estados-mentales llegó a parecer la pregunta correcta a plantear? Pero los efectos intelectuales de su persistencia parecen claros. Lo mismo que la distinción kantiana de sentido común entre las apariencias de las cosas y las cosas mismas llevó a los extremos del idealismo absoluto, así la persistencia de la pregunta de sentido comun «¿Bajo qué condiciones atribuiríamos estados mentales?» nos ha llevado al conductismo, al funcionalismo, a la IA fuerte, al materialismo eliminativo, a la postura intencional, y, sin duda, a otras confusiones conocidas sólo por los expertos.

En cuarto lugar: debido a nuestra concepción de la historia del incremento del conocimiento, hemos llegado a sufrir lo que Austin denominó «ivresse des grands profondeurs». No parece suficiente enunciar verdades humildes y obvias sobre la mente —queremos algo más profundo. Queremos un descubrimiento teórico. Y, desde luego, nuestro modelo de gran descubrimiento teórico proviene de la historia de las ciencias físicas. Soñamos con algún gran «estallido» en el estudio de la mente, esperamos y deseamos una ciencia cognitiva «madura». De este modo, el hecho de que los puntos de vista en cuestión sean implausibles y contraintuitivos no cuenta en contra de ellos. Por el contrario, puede incluso parecer un gran mérito del funcionalismo contemporáneo y de la inteligencia artificial que vayan totalmente en contra de nuestras in-

tuiciones. ¿Pues no es este el mismo rasgo que hace a las ciencias físicas tan deslumbrantes? Se ha mostrado que nuestras intuiciones ordinarias sobre el espacio y el tiempo o, también, sobre la solidez de la mesa que está ante nosotros son meras ilusiones reemplazadas por un conocimiento mucho más profundo del funcionamiento interno del universo. ¿No podría suceder que un gran estallido en el estudio de la mente mostrase, de manera similar, que las creencias que más firmemente mantenemos sobre nuestros estados mentales son igualmente ilusorias? ¿No es razonable que podamos esperar grandes descubrimientos que subviertan nuestras suposiciones de sentido común? Y, quién sabe, ¿no podría suceder que alguno de esos grandes descubrimientos fuera hecho por alguno de nosotros?

V. Socavar los cimientos

Un modo de enunciar algunas de las características más sobresalientes del argumento que estoy presentando es enunciarlas en oposición con los principios que he mencionado antes. Para hacer esto necesito, en primer lugar, hacer explícitas las distinciones entre *ontología*, *epistemología* y *causación*. Hay una distinción entre respuestas a las preguntas ¿Qué es? (ontología), ¿Cómo lo averiguamos? (epistemología) y ¿Qué lo hace? (causación). Por ejemplo, en el caso del corazón, la ontología es que es una extensa porción de tejido muscular que está en la cavidad torácica; la epistemología es que lo averiguamos usando estetoscopios, electrocardiogramas y, si estamos en un apuro, podemos abrir el tórax y echarle una mirada; y la causación es que el corazón bombea sangre a través del cuerpo. Teniendo presente esas distinciones, podemos empezar a trabajar.

1. *La conciencia tiene importancia.* Argumentaré que no hay manera de estudiar los fenómenos de la mente sin estudiar, implícita o explícitamente, la conciencia. La razón básica de esto es que no tenemos realmente noción alguna de lo mental aparte de nuestra noción de conciencia. Desde luego, en cualquier punto dado de la vida de una persona, la mayor parte de los fenómenos mentales de la existencia de esa persona no están presentes en la conciencia. En el modo formal, muchos de los predicados mentales que se me aplican en un instante dado tendrán condiciones de aplicación independientes de mis estados

conscientes en ese momento. Sin embargo, aunque la mayor parte de nuestra vida mental en cualquier punto dado es inconsciente, argumentaré que no tenemos concepción alguna de un estado mental inconsciente excepto en términos derivados de los estados mentales conscientes. Si estoy en lo correcto respecto de este asunto, entonces todas las discusiones recientes sobre estados mentales que, en principio, son inaccesibles a la conciencia son realmente incoherentes (sobre esta cuestión, véase el capítulo 7).

2. *No toda la realidad es objetiva; parte de ella es subjetiva.* Existe una persistente confusión entre la afirmación de que deberíamos intentar en todo lo posible eliminar los prejuicios subjetivos de la búsqueda de la verdad y la afirmación de que el mundo real no contiene elemento alguno que sea irreductiblemente subjetivo. Y esta confusión se basa, a su vez, en el sentido epistemológico de la distinción subjetivo/objetivo, y en el sentido ontológico. Epistémicamente, la distinción marca diferentes grados de independencia de las afirmaciones respecto de los caprichos de los valores especiales, prejuicios personales, puntos de vista y emociones. Ontológicamente, la distinción señala diferentes categorías de realidad empírica (sobre estas distinciones, véase el capítulo 4). Epistémicamente, el ideal de objetividad enuncia una meta valiosa aunque inalcanzable. Pero ontológicamente, la afirmación de que toda la realidad es objetiva es, neurológicamente, simple y llanamente falsa. En general, los estados mentales tienen una ontología irreductiblemente subjetiva, como tendremos ocasión de ver más adelante con algún detalle.

Si estoy en lo cierto al pensar que conciencia y subjetividad son esenciales para la mente, entonces la concepción de lo mental empleada por la tradición está mal concebida desde el principio, pues es, esencialmente, una concepción objetiva, de tercera persona. La tradición intenta estudiar la mente como si ésta consistiese en fenómenos neutrales, independientes de la conciencia y de la subjetividad. Pero tal enfoque deja fuera los rasgos cruciales que distinguen los fenómenos mentales de los no mentales. Y es esto más que cualquier otra razón lo que da cuenta de la implausibilidad de los puntos de vista que he mencionado al principio. Si se intenta tratar las creencias, por ejemplo, como fenómenos que no tienen conexión esencial alguna con la conciencia, entonces lo más probable es que uno se quede con la idea de que éstas sólo pueden definirse en términos de conducta externa (con-

ductismo), o en términos de relaciones de causa y efecto (funcionalismo), o de que no existen en absoluto (materialismo eliminativo), o de que el habla de creencias y deseos ha de interpretarse sólo como una cierta manera de hablar (la postura intencional). El último de los absurdos es intentar tratar la conciencia misma independientemente de la conciencia, esto es: tratarla solamente desde el punto de vista de la tercera persona, y esto lleva a la concepción de que la conciencia como tal, como eventos fenoménicos «internos», «privados», no existe realmente.

Algunas veces la tensión entre la metodología y lo absurdo de los resultados resulta visible. En la literatura reciente hay un debate sobre algo llamado «*qualia*» y el problema se supone que es: «¿Puede el funcionalismo dar cuenta de los *qualia*?». Lo que el problema revela es que la mente consta de *qualia*, por así decirlo, hasta el fondo. El funcionalismo no puede dar cuenta de los *qualia* porque está diseñado en torno a un tema distinto, a saber: las atribuciones de intencionalidad basadas en evidencia de tercera persona, mientras que los fenómenos mentales efectivos no tienen nada que ver con atribuciones, sino con la existencia de estados mentales conscientes e inconscientes, y ambos son fenómenos subjetivos, de primera persona.

3. *Puesto que es un error suponer que la ontología de lo mental es objetiva, es un error suponer que la metodología de una ciencia de la mente debe interesarse solamente por la conducta objetivamente observable.* Puesto que los fenómenos mentales están esencialmente conectados con la conciencia, y puesto que la conciencia es esencialmente subjetiva, se sigue que la ontología de lo mental es, esencialmente, una ontología de primera persona. Los estados mentales son siempre los estados mentales de alguien. Hay siempre una «primera persona», un «yo», que tiene esos estados mentales. La consecuencia de esto para la discusión presente es que el punto de vista de la primera persona es primario. En la práctica efectiva de la investigación queremos, desde luego, estudiar otras personas simplemente porque la mayor parte de nuestra investigación no es sobre nosotros mismos. Pero es importante subrayar que a lo que estamos intentando llegar cuando estudiamos otras personas es, precisamente, al punto de vista de la primera persona. Cuando lo estudiamos a *él* o a *ella*, lo que estamos estudiando es el *yo* que es *él* o *ella*. Esto no es un asunto epistémico.

A la luz de las distinciones entre ontología, epistemología y causa-

ción, si uno tuviera que resumir la crisis de la tradición en un párrafo, tal párrafo rezaría más o menos así:

La ontología subjetivista de lo mental parece intolerable. Parece intolerable metafísicamente que haya en el mundo entidades «privadas», irreductiblemente subjetivas, y epistemológicamente intolerable que exista una asimetría entre el modo en que cada persona conoce sus fenómenos mentales internos y el modo en que otros los conocen desde fuera. Esta crisis produce una huida de la subjetividad y la dirección de la huida es reescribir la *ontología* en términos de la *epistemología* y de la *causación*. Primero nos libramos de la subjetividad redefiniendo la ontología en términos de la tercera persona, bases epistémicas, conducta. Decimos: «Los estados mentales son sólo disposiciones de conducta» (conductismo), y cuando el absurdo se vuelve insoportable nos retiramos hacia la causación». Decimos: «Los estados mentales se definen por sus relaciones causales» (funcionalismo) o «Los estados mentales son estados computacionales» (IA fuerte).

La tradición supone, falsamente de acuerdo con mi punto de vista, que en el estudio de la mente uno está forzado a elegir entre «introspección» y «conducta». Hay diversos errores involucrados en esto; entre ellos están:

4. *Es un error suponer que sólo conocemos la existencia de los fenómenos mentales en los demás observando su conducta.* Creo que la «solución» tradicional al «problema de las otras mentes», aunque ha estado con nosotros durante siglos, no ha de sobrevivir ni siquiera un momento a una reflexión seria. Tengo más cosas que decir sobre estos problemas más adelante (en el capítulo 3), pero ahora diré sólo esto: si se piensa durante un momento en cómo sabemos que los perros y los gatos son conscientes, y que los ordenadores y los coches no lo son (y, dicho sea de paso, no hay duda alguna de que usted y yo sabemos esto), se verá que la base de nuestra certeza no reside en la «conducta», sino más bien en cierta concepción causal de cómo funciona el mundo. Puede verse que los perros y los gatos son, en ciertos aspectos relevantes, similares a nosotros. Tienen ojos, piel, orejas, etc. La «conducta» sólo tiene sentido como expresión o manifestación de una realidad mental subyacente, puesto que podemos ver las bases causales de lo mental y, por lo tanto, ver la conducta como una manifestación de lo mental. El principio de acuerdo con el cual «resolvemos» el problema de las otras mentes no es, voy a argumentar, el siguiente: misma-conducta-*ergo*-

mismos-fenómenos-mentales. Este es el viejo error que celosamente preserva el test de Turing. Si este principio fuese correcto, todos nosotros tendríamos que concluir que los aparatos de radio son conscientes puesto que exhiben conducta verbal inteligente. Pero no extraemos tal conclusión porque tenemos una «teoría» acerca de cómo funcionan los aparatos de radio. El principio de acuerdo con el cual «solucionamos el problema de las otras mentes» es el siguiente: mismas-causas-mismos-efectos, y causas-relevantemente-similares-efectos-relevantemente-similares. Por lo que concierne al conocimiento de otras mentes, la conducta no tiene interés alguno *por sí misma*; es más bien *la combinación de conducta con el conocimiento de los apoyos causales de la conducta* lo que forma las bases de nuestro conocimiento.

Pero incluso lo anterior me parece que hace demasiadas concesiones a la tradición, puesto que sugiere que nuestra postura básica hacia perros, gatos, aparatos de radio, y otras personas es epistémica; sugiere que en nuestras transacciones diarias con el mundo estamos ocupados en «resolver el problema de las otras mentes» y que los perros y los gatos pasan la prueba mientras que los aparatos de radio y los coches no lo logran. Pero esta sugerencia es errónea. Excepto en casos extraños, no resolvemos el problema de las otras mentes, porque no se plantea. Nuestras capacidades de Trasfondo para habérnoslas con el mundo nos capacitan para tratar con la gente de una manera y con los coches de otra, pero no generamos, además, una hipótesis al efecto de que esta persona es consciente y que el coche no lo es, excepto en los casos inusuales. Diré más cosas sobre esto más adelante (en los capítulos 3 y 8).

Ciertamente, en las ciencias surgen cuestiones epistémicas, pero las cuestiones epistémicas no son más esenciales para comprender la naturaleza de la mente de lo que lo son para comprender la naturaleza de los fenómenos estudiados en cualquier otra disciplina. ¿Por qué habrían de serlo? Hay interesantes cuestiones epistémicas en historia, por ejemplo, sobre el conocimiento del pasado, o sobre el conocimiento de entidades inobservadas en física. Pero la cuestión «¿*Cómo* ha de verificarse la existencia de los fenómenos?» no debe de confundirse con la cuestión «¿*Cuál* es la naturaleza de los fenómenos cuya existencia se verifica?». La cuestión crucial no es «¿Bajo qué condiciones deberíamos *atribuir* estados mentales a otras personas?», sino más bien «¿Qué es lo que la gente *tiene efectivamente* cuando tiene estados mentales?». «¿Qué son los estados mentales» como pregunta distinta de «¿Cómo averiguamos

cosas sobre ellos y cómo funcionan causalmente en la vida de un orga nismo?».

No quiero que este punto se entienda mal: no estoy diciendo qu sea fácil averiguar cosas sobre los estados mentales, y que no tengamo que preocuparnos de cuestiones epistémicas. Este no es el asunto. Pien so que es inmensamente difícil estudiar los fenómenos mentales, y l única guía metodológica es la guía universal: usa cualquier instrumen to o arma que tengas a mano, y remata con cualquier instrumento (arma que funcione. Lo que quiero decir aquí es diferente: la epistemo logía del estudio de lo mental no determina su ontología en mayor me dida que la epistemología de cualquier otra disciplina determina su pro pia ontología. Por el contrario, en el estudio de la mente, como e cualquier otro estudio, el objetivo total de la epistemología es llegar a la ontología preexistente.

5. *La conducta o las relaciones causales con la conducta no so esenciales para la existencia de fenómenos mentales.* Creo que la rela ción de los estados mentales con la conducta es puramente contingen te. Es fácil ver esto cuando consideramos cómo es posible tener los es tados mentales sin la conducta, y la conducta sin los estados mentale (daré algunos ejemplos en el capítulo 3). Sabemos que, causalmente los procesos cerebrales son suficientes para cualquier estado mental y que la conexión entre esos procesos cerebrales y el sistema nervioso motor es una conexión neurofisiológica contingente como cualquie otra.

6. *Es inconsistente con lo que sabemos sobre el universo y sobre nuestro lugar en él suponer que todo es cognoscible por nosotros.* Nuestros cerebros son los productos de ciertos procesos de evolución y, como tales, son simplemente los más desarrollados en toda una serie de caminos de evolución que incluyen los cerebros de los perros, babuinos, delfines, etc. Ahora bien, nadie supone que podamos hacer que los perros entiendan mecánica cuántica; el cerebro del perro simplemente no está desarrollado hasta ese punto. Y es fácil imaginar un ser que esté más desarrollado en la misma línea de progreso evolutivo en que nosotros estamos y que sea para nosotros, más o menos, lo que nosotros somos para los perros. Lo mismo que pensamos que los perros no pueden entender la mecánica cuántica, este producto imaginario de la evolución concluiría que, aunque los humanos pueden entender la mecáni-

ca cuántica, hay todavía una gran cantidad de cosas que el cerebro humano no puede captar.[12] Es una buena idea preguntarnos a nosotros mismos: ¿qué pensamos que somos? Y, al menos, parte de la respuesta es que somos bestias biológicas seleccionadas para enfrentarnos con entornos cazadores-recolectores y, por lo que sabemos, no hemos sufrido ningún cambio significativo en nuestra dotación genética en varios miles de años. Afortunadamente (o desafortunadamente) la naturaleza es despilfarradora, y así como un solo macho produce esperma suficiente para repoblar la tierra, del mismo modo nosotros tenemos muchas más neuronas de las que necesitamos para una existencia cazadora-recolectora. Creo que el fenómeno del exceso de neuronas —como distinto de, digamos, los pulgares antagónicos— es la clave para comprender cómo salimos del estadio cazador-recolector y producimos filosofía, ciencia, tecnología, neurosis, publicidad, etc. Pero no deberíamos olvidar nunca quiénes somos; y para seres tales como nosotros, es un error suponer que todo lo que existe es comprensible para nuestros cerebros. Desde luego, metodológicamente podemos actuar como si pudiésemos entenderlo todo, puesto que no hay manera de conocer lo que no podemos conocer: para conocer los límites del conocimiento deberíamos de conocer ambos lados del límite. Así pues, la omnisciencia potencial es aceptable como recurso heurístico, pero sería autoengañarnos suponer que se trata de un hecho.

Además, sabemos que muchos seres de nuestra tierra tienen estructuras neurofisiológicas que son suficientemente diferentes de las nuestras de modo que nos puede ser literalmente incognoscible a qué se parecen las experiencias de esos seres. Discutiré un ejemplo de esto en el capítulo 3.

7. *La concepción cartesiana de lo físico, la concepción de la realidad física como* res extensa, *simplemente no es adecuada para describir los hechos que corresponden a enunciados sobre la realidad física.* Cuando llegamos a la proposición de que la realidad es física, llegamos a lo que es quizás el punto crucial de toda la discusión. Cuando pensamos en lo «físico», pensamos en cosas tales como moléculas y átomos y partículas subatómicas. Y pensamos que son físicas en un sentido que se opone a lo mental, y que cosas como sensaciones y do-

12. Una observación parecida a esta la hace Noam Chomsky (1975).

lor son mentales. Y si hemos crecido en nuestra cultura, pensamos también que esas dos categorías tienen que agotar todo lo que existe. Pero la pobreza de esas categorías se vuelve aparente tan pronto como se empieza a pensar sobre los diferentes géneros de cosas que contiene el mundo, esto es: tan pronto como se empieza a pensar sobre los hechos que corresponden a las distintas clases de enunciados empíricos. Así pues, si uno piensa sobre los problemas de los balances de pagos, las oraciones no gramaticales, las razones para mirar con suspicacia a la lógica modal, mi destreza para esquiar, el gobierno del estado de California y los goles marcados en los partidos de fútbol, uno está menos inclinado a pensar que todo debe de categorizarse como mental o como físico. De la lista que acabo de dar, ¿qué cosas son mentales y qué cosas son físicas?

Hay al menos tres cosas erróneas en nuestra concepción tradicional de que la realidad es física. En primer lugar, como he señalado, la terminología está diseñada en torno a una falsa oposición entre lo «físico» y lo «mental» y, como he afirmado ya, esto es un error. En segundo lugar, si pensamos en lo físico en términos cartesianos como *res extensa* entonces resulta obsoleto suponer, incluso en términos de física, que la realidad física lo es de acuerdo con esta definición. Desde la teoría de la relatividad pensamos en, por ejemplo, los electrones como puntos de masa-energía. En tercer lugar (y más importante para la discusión presente): es un error muy acusado el suponer que la cuestión crucial para la ontología es «¿Qué clase de cosas existen en el mundo?» como opuesta a «¿Qué tiene que ser el caso en el mundo para que nuestros enunciados empíricos sean verdaderos?».

Noam Chomsky dijo una vez (en una conversación) que tan pronto como llegamos a entender algo lo llamamos «físico». De acuerdo con este punto de vista, trivialmente, cualquier cosa o es física o es ininteligible. Si pensamos en la constitución del mundo entonces, desde luego, el mundo está hecho de partículas, y las partículas están entre nuestros paradigmas de lo físico. Y si hemos de llamar físico a todo aquello que está hecho de partículas físicas entonces, trivialmente, todo lo que hay en el mundo es físico. Pero decir esto no equivale a negar que el mundo contenga goles marcados en partidos de fútbol, tipos de interés, gobiernos y dolores. Todas estas cosas tienen su propio modo de existir —deportivo, económico, político, mental, etc.

La conclusión es esta: una vez que se ve la incoherencia del dualismo, se puede ver también que el monismo y el materialismo están

igual de equivocados. Los dualistas preguntaban: «¿Cuántos géneros de cosas y de propiedades hay?» y contaban hasta dos. Los monistas, enfrentados a la misma pregunta, sólo llegaban hasta uno. Pero el error real era empezar a contar. El monismo y el materialismo se definen en términos de dualismo y mentalismo, y puesto que las definiciones de dualismo y mentalismo son incoherentes, monismo y materialismo heredan esa incoherencia. Es usual pensar que el dualismo se presenta en dos sabores, dualismo de substancias y dualismo de propiedades; pero a estos yo quiero añadir un tercero, que llamaré «dualismo conceptual». Este punto de vista consiste en tomar muy en serio los conceptos dualistas, esto es: consiste en el punto de vista de que, en algún sentido importante, «físico» implica «no mental» y «mental» implica «no físico». Tanto el dualismo tradicional como el materialismo presuponen el dualismo conceptual así definido. Introduzco esta definición para aclarar por qué me parece mejor pensar en el materialismo como siendo, realmente, una forma de dualismo. Se trata de aquella forma de dualismo que comienza aceptando las categorías cartesianas. Creo que si se toman seriamente esas categorías —las categorías de mental y físico, mente y cuerpo— como un dualismo consistente, uno se verá eventualmente forzado a abrazar el materialismo. El materialismo es entonces, en un sentido, la flor más delicada del dualismo. Paso ahora a exponer sus dificultades e historia reciente.

2. LA HISTORIA RECIENTE DEL MATERIALISMO: EL MISMO ERROR UNA Y OTRA VEZ

I. El misterio del materialismo

¿A qué se supone que equivale exactamente la doctrina conocid
como «materialismo»? Podría pensarse que tal doctrina consiste en e
punto de vista de que la microestructura del mundo está formada ente
ramente por partículas materiales. La dificultad, sin embargo, es qu
este punto de vista es consistente con casi cualquier filosofía de la men
te, excepto posiblemente el punto de vista cartesiano de que además d
las partículas físicas hay almas «inmateriales» o substancias mentales
entidades espirituales que sobreviven a la destrucción de nuestros cuer
pos y continúan viviendo inmortalmente. Pero hoy en día, por lo que sé
nadie cree en la existencia de substancias espirituales e inmortales, ex
cepto teniendo como base creencias religiosas. Por lo que sé, no ha
motivaciones puramente filosóficas o científicas para aceptar la exis
tencia de substancias mentales inmortales. Así pues, dejando de lado l
oposición a la creencia en almas inmortales motivada religiosamente
queda aún la cuestión siguiente: ¿a qué se supone que equivale exacta
mente el materialismo en filosofía de la mente? ¿A qué puntos de vist
hemos de suponer que se opone?

Si se leen las obras tempranas de nuestros contemporáneos que s
describen a sí mismos como materialistas —J. J. C. Smart (1965), U. T
Place (1956) y D. Armstrong (1968), por ejemplo— parece claro qu
cuando aseveran la identidad de lo mental con lo físico están afirman
do algo más que la simple negación de la existencia en el mundo d
cualquier fenómeno mental irreductible. Quieren negar que exista
cualesquiera propiedades fenomenológicas irreductibles tales como l
conciencia o los *qualia*. Ahora bien, ¿por qué tienen tanto afán en ne

:ar la existencia de fenómenos mentales intrínsecos irreductibles? ¿Por
ué no conceden simplemente que esas propiedades son propiedades
iológicas ordinarias de nivel superior de sistemas neurofisiológicos
ıles como los cerebros humanos?

Pienso que la respuesta a esto es extremadamente compleja, pero al
ıenos parte de la respuesta tiene que ver con el hecho de que aceptan
ıs categorías cartesianas tradicionales y, junto con las categorías, el
'ocabulario que las acompaña y sus implicaciones. Pienso que dar por
entado desde este punto de vista la existencia e irreductibilidad de los
enómenos mentales sería equivalente a dar por sentado algún género
le cartesianismo. En sus términos, se trataría de un «dualismo de pro-
ıiedades» más bien que de un «dualismo de substancias», pero desde
u punto de vista, el dualismo de propiedades sería tan inconsistente
:on el materialismo como el dualismo de substancias. Resultará obvio
ıhora que me opongo a las suposiciones que subyacen en sus puntos de
'ista. Aquello en lo que quiero insistir, una y otra vez, es que se pueden
ıceptar los hechos obvios de la física —por ejemplo, que el mundo está
'ormado enteramente por partículas físicas en campos de fuerza— sin
ıegar al mismo tiempo los hechos obvios sobre nuestras experiencias
—por ejemplo, que todos nosotros somos conscientes y que nuestros
:stados conscientes tienen propiedades fenomenológicas específicas
:ompletamente *irreductibles*. El error consiste en suponer que esas dos
esis son inconsistentes, y ese error se deriva de aceptar las presuposi-
:iones que subyacen en el vocabulario tradicional. Mi punto de vista,
ɿuiero subrayarlo, no es una forma de dualismo. Rechazo tanto el dua-
ismo de propiedades como el dualismo de substancias; pero precisa-
nente por las mismas razones por las que rechazo el dualismo, rechazo
ambién el materialismo y el monismo. El gran error es suponer que se
lebe elegir entre esos dos puntos de vista.

El no lograr ver la consistencia del mentalismo ingenuo con el fisi-
:alismo ingenuo es lo que lleva a esas discusiones tan problemáticas en
a historia primitiva de este asunto; en ellas los autores tratan de en-
:ontrar un vocabulario «neutral respecto al tema» o de evitar lo que lla-
nan «colgantes nomológicos» (*nomological danglers*) (Smart, 1965).
Téngase en cuenta que nadie siente la necesidad de que, pongamos por
:aso, la digestión tenga que describirse en un vocabulario «neutral res-
pecto al tema». Nadie siente el impulso de decir: «Hay algo que está su-
:ediendo dentro de mí y que es parecido a lo que sucede cuando digie-
:o pizza». Sin embargo, sí sienten el impulso de decir: «Hay algo que

está sucediendo en mí que se parece a lo que sucede cuando veo una naranja». El impulso es intentar encontrar una descripción de los fenómenos que no use el vocabulario mentalista. Pero ¿para qué se hace esto? Los hechos siguen siendo los mismos. El hecho es que los fenómenos mentales tienen propiedades mentalistas, lo mismo que lo que está sucediendo en mi estómago tiene propiedades digestivas. No nos libramos de esas propiedades simplemente encontrando un vocabulario alternativo. Los filósofos materialistas quieren negar la existencia de propiedades mentales sin negar la realidad de *algunos* fenómenos que subyacen en el uso de nuestro vocabulario mentalista. Así pues, tienen que encontrar un vocabulario alternativo para describir los fenómenos.[1] Pero de acuerdo con mi explicación, todo esto es una pérdida de tiempo. Deberían darse por sentados los fenómenos mentales (y, por lo tanto, físicos) para empezar, de la misma manera que uno da por sentados los fenómenos digestivos en el estómago.

En este capítulo quiero examinar, más bien brevemente, la historia del materialismo durante el último medio siglo. Creo que esta historia exhibe un modelo de argumento y contraargumento más bien problemático pero muy revelador, que ha operado en la filosofía de la mente desde el positivismo de los años treinta. Este modelo no es siempre visible en la superficie. Ni es siquiera visible en la superficie que se está hablando de los mismos problemas. Pero, contrariamente a las apariencias superficiales, ha habido realmente un tema de discusión principal en la filosofía de la mente durante, más o menos, los últimos cincuenta años y este es el problema mente-cuerpo. A menudo los filósofos pretenden hablar sobre algo distinto —tal como, por ejemplo, el análisis de la creencia o la naturaleza de la conciencia— pero, casi invariablemente, sale a la superficie que no están interesados en rasgos especiales de la creencia o de la conciencia. No están interesados en cómo el creer difiere del suponer o del hacer hipótesis, sino que más bien lo que quie-

1. Un buen ejemplo de esto se encuentra en Richard Rorty (1979). Nos pide que imaginemos una tribu que no dice «Tengo dolor», sino más bien «Mis fibras C están siendo estimuladas». Bien, imaginémonos un caso semejante. Imaginémonos una tribu que no quiere usar nuestro vocabulario mentalista. ¿Qué se sigue de ello? O tienen dolores como los tenemos nosotros o no los tienen. Si los tienen, entonces el hecho de que rehúsen llamarlos dolores no tiene interés. Los hechos siguen siendo los mismos independientemente de cómo nosotros o ellos elijamos describirlos. Si, por otra parte, no tienen realmente dolor alguno, entonces son completamente diferentes de nosotros y su situación no tiene relevancia alguna para la realidad de nuestros fenómenos mentales.

en es poner a prueba sus convicciones sobre el problema mente-cuerpo con el *ejemplo* de la creencia. Lo mismo sucede con la conciencia: sorprendentemente, hay muy poca discusión sobre la conciencia como tal; los materialistas ven la conciencia más bien como un «problema» especial para una teoría materialista de la mente. Esto es: quieren encontrar un modo de «manejar» la conciencia, dado su materialismo.[2]

El patrón que esas discusiones parecen seguir casi invariablemente es este. Un filósofo avanza una teoría materialista de la mente. Hace esto desde la firme convicción de que alguna versión de la teoría materialista de la mente tiene que ser la correcta —después de todo, ¿no sabemos por los descubrimientos de la ciencia que en el universo no hay otra cosa que partículas físicas y campos de fuerza que actúan sobre las partículas físicas? Y seguramente debe de ser posible proporcionar una explicación de los seres humanos que sea consistente y coherente, de manera general, con nuestra explicación de la naturaleza. Y seguramente, ¿no se sigue de esto que nuestra explicación de los seres humanos debe ser materialismo puro? Así pues, el filósofo se plantea dar una explicación materialista de la mente. Encuentra entonces dificultades. Siempre parece que está dejando algo fuera. El patrón general de discusión es que las críticas de la teoría materialista usualmente tienen una forma más o menos técnica, pero, de hecho, de manera subyacente a las objeciones técnicas, hay una objeción mucho más profunda, y esa objeción más profunda puede enunciarse muy simplemente: la teoría en cuestión ha dejado fuera la mente; ha dejado fuera algún rasgo esencial de la mente, tal como la conciencia o los *qualia* o el contenido semántico. Este patrón puede verse una y otra vez. Se avanza una tesis materialista. Pero la tesis encuentra dificultades; las dificultades toman formas diferentes, pero son siempre manifestaciones de una dificultad subyacente más profunda, a saber: la tesis en cuestión niega hechos obvios que todos conocemos sobre nuestras mentes. Y esto lleva a esfuerzos cada vez más frenéticos para mantenerse en las tesis materialistas e intentar derrotar los argumentos avanzados por aquellos que

2. Resulta un hecho interesante el que en tres libros recientes en los que la palabra «conciencia» aparece en sus títulos —el de Paul Churchland *Matter and Consciousness* (1984), el de Ray Jackendoff *Consciousness and the Computational Mind* (1987) y el de William Lycan *Consciousness* (1987)— haya poco o ningún esfuerzo para dar una explicación, o una teoría, de la conciencia. La conciencia no es un tema que se considere como algo que merece la pena tratar por sí mismo, sino como un problema incordiante para la filosofía de la mente materialista.

insisten en preservar los hechos. Después de algunos años de maniobras desesperadas para dar cuenta de las dificultades, se plantea algún nuevo desarrollo que se pretende que resuelva las dificultades, pero entonces nos topamos con que tal desarrollo encuentra nuevas dificultades, sólo que las nuevas dificultades no son tan nuevas —son realmente las mismas dificultades viejas.

Si concibiéramos la filosofía de la mente de los últimos cincuenta años como si se tratara de un individuo, diríamos que nos hallamos ante una persona que es un neurótico compulsivo, cuya neurosis adopta la forma de repetir una y otra vez el mismo patrón de conducta. Tengo la experiencia de que la neurosis no puede ser curada por medio de un ataque frontal. No basta con señalar los errores lógicos que se cometen. La refutación directa no conduce más que a la repetición del mismo patrón de conducta neurótica. Lo que tenemos que hacer es ir detrás de los síntomas y descubrir, en primer lugar, los supuestos inconscientes que llevaron a la conducta neurótica. Estoy convencido, después de discutir estos problemas durante algunos años, de que todas las partes en discordia, con algunas excepciones, son prisioneras de cierto conjunto de categorías verbales. Son prisioneras de cierta terminología, una terminología que se remonta, como mínimo, a Descartes y, para poder vencer a la conducta compulsiva, deberemos examinar los orígenes inconscientes de las diferencias. Deberemos descubrir qué es lo que cada cual da por sentado para que la discusión se mantenga y se reproduzca interminablemente.

No desearía que mi uso de una analogía terapéutica se considerase que implica un apoyo general a los modos psicoanalíticos de explicación en asuntos intelectuales. De modo que alteraremos del siguiente modo la metáfora psicoanalítica: quiero sugerir que mi actual tarea es un poco similar a la de un antropólogo que trata de describir la conducta exótica de una tribu lejana. La tribu tiene un conjunto de patrones conductuales y una metafísica que debemos tratar de desvelar y comprender. Es fácil burlarse de los trucos de la tribu de los filósofos de la mente y, debo confesarlo, no siempre he resistido la tentación de actuar de ese modo. Pero, al menos al principio, debo insistir en que la tribu somos nosotros —nosotros somos los poseedores de los supuestos metafísicos que hacen posible la conducta de la tribu. De modo que, antes de presentar un análisis y una crítica de la conducta de la tribu, quiero presentar una idea que todos nosotros debemos considerar aceptable dado que la idea es realmente parte de nuestra cultura científica con

emporánea. Y, sin embargo, voy a argumentar más adelante que la dea es incoherente; sólo es otro síntoma del mismo patrón neurótico.

He aquí la idea. Pensamos que la cuestión siguiente debe tener sen-ido: ¿Cómo es posible que fragmentos no inteligentes de materia pro-luzcan inteligencia? ¿Cómo es posible que los fragmentos no inteligen-es de materia de nuestro cerebro produzcan la conducta inteligente en la ןue todos nosotros estamos implicados? Esta nos parece una pregunta erfectamente inteligible. En realidad, parece un valioso proyecto de in-estigación, y, de hecho, es un proyecto de investigación que tiene mu-hos seguidores[3] y, algunas veces, muchos fondos económicos.

Porque encontramos inteligible la cuestión anterior, encontramos lausible la siguiente respuesta a ella. Los fragmentos no inteligentes le materia pueden producir inteligencia a causa de su *organización.* Los fragmentos no inteligentes de materia están *organizados* de ciertas naneras dinámicas, y es la organización dinámica la que es constituti-a de la inteligencia. De hecho, podemos reproducir artificialmente la orma de organización dinámica que hace posible la inteligencia. La es-ructura subyacente en esa organización se denomina un «ordenador»; l proyecto de programar el ordenador se denomina «inteligencia arti-icial»; y el ordenador produce inteligencia cuando está operando por-ןue está implementando el programa de ordenador adecuado con los *nputs* y *outputs* adecuados.

¿Encuentra el lector verosímil al menos esta respuesta? Debo con-esar que es posible conseguir que a mí me parezca muy verosímil, e in-luso pienso que aquel al que no le suene ni siquiera remotamente ve-osímil no es probable que sea un miembro completamente socializado le nuestra cultura intelectual contemporánea. Más adelante, trataré de nostrar que tanto la pregunta como la respuesta son incoherentes. Cuan-lo planteamos la pregunta y le damos respuesta en estos términos, no enemos, en realidad, la más mínima idea de qué estamos hablando. Pero presento el ejemplo aquí porque quiero que parezca natural, in-luso prometedora, como proyecto de investigación.

Unos pocos párrafos más atrás he dicho que la historia del materia-ismo filosófico en el siglo XX exhibe un patrón curioso, un patrón en el ןue hay una tensión recurrente entre, por una parte, el ansia materialis-

3. En su recensión del libro de Marvin Minsky *Society of Mind*, Bernard Williams 1987) escribe: «Lo que está en cuestión en esta [I.A.] investigación es, precisamente, si os sistemas inteligentes pueden estar compuestos de materia no inteligente».

ta por proporcionar un análisis de los fenómenos mentales que no haga referencia a nada intrínseca o irreductiblemente mental y, por otra parte, el requisito intelectual general que acepta todo investigador de no decir nada que sea obviamente falso. Para dejar que este patrón se muestre por sí mismo, deseo dar un breve repaso, tan neutral y objetivamente como soy capaz, del patrón de tesis y respuestas que ha ejemplificado el materialismo. El propósito de lo que sigue es proporcionar evidencia para las afirmaciones que se hacen en el capítulo 1, dando ejemplos de las tendencias que he identificado.

II. CONDUCTISMO

Al principio era el conductismo. El conductismo se presentaba en dos variedades: «conductismo metodológico» y «conductismo lógico». El conductismo metodológico es una estrategia de investigación en psicología, con la propuesta de que la ciencia psicológica debe consistir en el descubrimiento de las relaciones entre los *inputs* estimulativos y los *outputs* conductuales (Watson, 1925). Una ciencia empírica rigurosa, de acuerdo con este punto de vista, no hace referencia alguna a elementos introspectivos misteriosos o mentalistas.

El conductismo lógico da incluso un paso más e insiste en que no existen elementos tales a los que referirse, excepto en la medida en que existen como forma de conducta. De acuerdo con el conductismo lógico, es un asunto de definición, un asunto de análisis lógico, el que los términos mentales puedan definirse en términos de conducta, el que las oraciones sobre la mente puedan traducirse en términos de oraciones sobre conducta, sin ningún tipo de residuo (Hempel, 1949; Ryle, 1949). De acuerdo con el conductista lógico, muchas de las oraciones así traducidas serán de forma hipotética, porque los fenómenos mentales en cuestión no consisten en que se den realmente ciertos fenómenos, sino, más bien, en ciertas disposiciones a la conducta. Así, de acuerdo con el análisis conductista habitual, decir que Juan cree que va a llover es decir sólo que Juan estará dispuesto a cerrar las ventanas, guardar las herramientas del jardín y coger el paraguas si sale a la calle. En el modo material de habla, el conductismo pretende que la mente es sólo conducta y disposiciones a comportarse. En el modo formal, consiste en el punto de vista de que las oraciones sobre los fenómenos mentales pueden traducirse a oraciones sobre la conducta real o posible.

Las objeciones al conductismo pueden dividirse en dos clases: las objeciones de sentido común y las que son más o menos técnicas. Una objeción de sentido común obvia es la de que el conductista deja de lado los fenómenos mentales en cuestión. El análisis conductista no deja nada para la experiencia subjetiva de pensar o sentir; sólo tiene en cuenta los patrones de conducta objetivamente observables.

El conductismo lógico ha tenido que enfrentarse a algunas objeciones más o menos técnicas. En primer lugar, los conductistas nunca lograron aclarar completamente la noción de «disposición». Nadie ha podido proporcionar una explicación satisfactoria de qué tipos de antecedentes deberían incorporarse a los enunciados hipotéticos para producir un análisis disposicional adecuado de los términos mentales en términos conductuales (Hampshire, 1950; Geach, 1957). En segundo lugar, parecía haber un problema consistente en cierto tipo de circularidad en el análisis: para dar un análisis de la creencia en términos de conducta, parece que hay que referirse al deseo; para analizar el deseo, hay que referirse a la creencia (Chisholm, 1957). Por considerar el ejemplo anterior, tratamos de analizar la hipótesis de que Juan cree que lloverá en términos de la hipótesis de que, si las ventanas están abiertas, Juan las cerrará, y otras semejantes. Queremos analizar el enunciado categórico de que Juan cree que va a llover en términos de ciertos enunciados hipotéticos sobre lo que hará Juan en determinadas condiciones. Sin embargo, la creencia de Juan de que va a llover sólo se manifestará por medio de la conducta de cerrar las ventanas si damos por supuestas hipótesis adicionales como la de que Juan no quiere que el agua de la lluvia se cuele por las ventanas, y la de que Juan cree que el agua puede entrar por las ventanas abiertas. Si no hay nada que desee más que el que el agua entre a raudales por sus ventanas no tendrá la disposición a cerrarlas. Sin ese tipo de hipótesis sobre los deseos de Juan (y sus creencias restantes), parece que no podemos comenzar a analizar ninguna oración sobre sus creencias originales. Pueden hacerse observaciones similares respecto al análisis de los deseos. Parece que tales análisis requieren la referencia a las creencias.

Una tercera objeción técnica al conductismo era la de que dejaba de lado las relaciones causales entre los estados mentales y la conducta (Lewis, 1966). Por ejemplo, al identificar el dolor con la disposición a la conducta de dolor, el conductismo deja de lado el hecho de que el dolor *causa* la conducta de dolor. Del mismo modo, si tratamos de *anali-*

zar las creencias y los deseos en términos de conducta, ya no podemos decir que las creencias y los deseos *causan* la conducta.

Aunque quizás la mayoría de las polémicas en las publicaciones filosóficas tratan de las objeciones «técnicas», de hecho, son las objeciones de sentido común las más embarazosas. El absurdo del conductismo radica en el hecho de que niega la existencia de los estados mentales internos como algo adicional a la conducta externa (Ogden y Richards, 1926). Y, como sabemos, esto va frontalmente en contra de nuestras experiencias ordinarias de lo que se siente siendo un ser humano. Por esta razón, los conductistas fueron irónicamente acusados de «simular la anestesia»[4] y fueron el objetivo de gran número de bromas bastante malas (por ejemplo, conductista número 1 a conductista número 2 después de hacer el amor: «Fue estupendo para ti, ¿qué tal me fue a mí?»). Esta objeción de sentido común al conductismo se expresó en forma de argumentos que apelaban a nuestras intuiciones. Una de ellas es la objeción del superactor-superespartano (Putnam, 1963). Es fácil imaginar un actor de extraordinaria técnica que pudiera imitar perfectamente la conducta de alguien que tuviera dolor aunque no sintiera ningún dolor, y es también posible imaginar un superespartano que pudiera soportar el dolor sin dar ninguna señal de estar sintiéndolo.

III. Teorías de la identidad de tipos

El conductismo lógico pretendía ser una verdad analítica. Aseveraba una conexión definicional entre los conceptos mentales y los conductuales. En la historia reciente de las filosofías materialistas de la mente, fue sustituido por la «teoría de la identidad», que pretendía que, como asunto de hecho empírico, contingente, sintético, los estados mentales eran idénticos a los estados del cerebro y del sistema nervioso central (Place, 1956; Smart, 1965). De acuerdo con los teóricos de la identidad, no había ningún absurdo en el supuesto de que pudiera haber fenómenos mentales separados, independientes de la realidad material; sólo que, como asunto de hecho, nuestros estados mentales como los dolores eran idénticos a ciertos estados de nuestro sistema nervioso. En

4. No conozco el origen de esta expresión, pero deriva probablemente de la caracterización que Ogden y Richards hicieron de Watson como «que exhibía anestesia general» (1926, p. 23 de la edición de 1949).

este caso, se pretendía que los dolores eran idénticos a estimulaciones de las fibras-C.[5] Descartes *podría* haber estado en lo cierto al pensar que había fenómenos mentales separados; sólo que, *como asunto de hecho*, resultó que estaba equivocado. Los fenómenos mentales no eran nada más que estados del cerebro y del sistema nervioso central. Se suponía que la identidad entre la mente y el cerebro era una identidad empírica, como se suponía de la identidad entre el relampagueo y las descargas eléctricas (Smart, 1965), o entre el agua y las moléculas de H_2O (Feigl, 1958; Shaffer, 1961), eran identidades empíricas y contingentes. Sucedió que resultó ser un descubrimiento científico que los fogonazos no eran sino corrientes de electrones y que el agua, en sus distintas formas, no era sino conjuntos de moléculas de H_2O.

Como en el caso del conductismo, podemos dividir las dificultades de la teoría de la identidad en objeciones «técnicas» y objeciones de sentido común. En este caso, la objeción de sentido común adopta la forma de un dilema. Supongamos que la teoría de la identidad es, como pretenden sus defensores, una verdad empírica. Si lo es, entonces debe haber rasgos del fenómeno en cuestión que sean lógicamente independientes para permitir su identificación en la parte de la derecha del enunciado de identidad de un modo diferente al que es utilizado para su identificación en la parte izquierda (Stevenson, 1960). Si, por ejemplo, los dolores fueran idénticos a sucesos neurofisiológicos, debería haber dos conjuntos de rasgos, rasgos dolorosos y rasgos neurofisiológicos, de tal modo que esos dos conjuntos de rasgos nos permitan fijar ambos lados del enunciado sintético de identidad. Así, supongamos, por ejemplo, que tenemos un enunciado de la forma:

El suceso doloroso x es idéntico al suceso neurofisiológico y.

5. Menciono esta forma de hablar de las «fibras-C» con cierta intranquilidad porque toda esta discusión muestra cierta carencia de información. Independientemente de los méritos y deméritos del materialismo, por motivos estrictamente neurofisiológicos, no puede defenderse que las fibras-C deban ser el *locus* de las sensaciones de dolor. Las fibras-C son un tipo de axón que transmite ciertos tipos de señales dolorosas desde las terminales nerviosas periféricas hasta el sistema nervioso central. Otras señales dolorosas son transmitidas por las fibras A-Delta. Las fibras-C funcionan como caminos de tránsito para que los estímulos lleguen al cerebro donde realmente sucede todo. Hasta donde sabemos, los sucesos neurofisiológicos que son responsables de las sensaciones de dolor se dan en el tálamo, el sistema límbico, el córtex somático-sensorial y, es posible, otras regiones. (Para estas cuestiones, véase cualquier libro de texto.)

Entendemos tal enunciado porque comprendemos que uno y el mismo suceso ha sido identificado por medio de dos tipos diferentes de propiedades, propiedades de dolor y propiedades neurofisiológicas. Pero, si esto es así, parece que nos enfrentamos a un dilema: o bien los rasgos dolorosos son subjetivos, mentales e introspectivos, o no lo son. Si lo son, no hemos conseguido desembarazarnos de la mente, después de todo. Todavía nos las tenemos que ver con una forma de dualismo, por más que sea un dualismo de propiedades en vez de un dualismo de substancias. Nos las tenemos que ver todavía con conjuntos de propiedades mentales, aunque nos hayamos desembarazado de las substancias mentales. Si, por otra parte, tratamos de considerar que «doloroso» no nombra un rasgo mental subjetivo de ciertos sucesos neurofisiológicos, su significado se nos presenta como algo completamente misterioso y sin explicación alguna. Como en el caso del conductismo, hemos dejado de lado a la mente. Ahora, ya no tenemos ningún medio para especificar esos rasgos mentales y subjetivos de nuestras experiencias.

Espero que esté claro que esta no es más que una repetición de la objeción de sentido común al conductismo. En el caso que nos ocupa, la hemos planteado bajo la forma de un dilema: o el materialismo de la identidad deja a un lado la mente o no lo hace. Si lo hace, es falso. Si no lo hace, no es materialismo.

Los teóricos de la identidad australianos pensaron que tenían una respuesta a esta objeción. La respuesta era intentar describir los denominados rasgos mentales en un vocabulario «neutral respecto al tema». La idea era la de obtener una descripción de los rasgos mentales que no mencionara el hecho de que eran mentales (Smart, 1965). Ciertamente, podemos hacer tal cosa: es posible mencionar dolores sin mencionar el hecho de que son dolores, del mismo modo que es posible mencionar aeroplanos sin mencionar el hecho de que son aeroplanos. Es decir, uno puede mencionar un aeroplano diciendo «una cierta instancia de propiedad perteneciente a las United Airlines», y uno puede referirse a una postimagen de amarillo-naranja diciendo «cierto suceso que me pasa a mí y que es semejante a los sucesos que me pasan cuando veo una naranja». Pero el hecho de que sea posible mencionar un fenómeno sin especificar sus características esenciales no quiere decir que no exista y que no tenga esas características esenciales. Todavía se trata de dolor o de una postimagen, o de un aeroplano, aun cuando nuestras descripciones no lleguen a mencionar esos hechos.

Otra objeción más «técnica» a la teoría de la identidad fue la si-

guiente: parece improbable que para cada tipo de estado mental haya un, y sólo uno, tipo de estado neurofisiológico al que sea idéntico. Incluso si mi creencia de que Denver es la capital de Colorado es idéntica a cierto estado de mi cerebro, parece excesivo esperar que cualquiera que crea que Denver es la capital de Colorado haya de poseer una configuración neurofisiológica idéntica en su cerebro (Block y Fodor, 1972; Putnam, 1967). Y, a través de las diversas especies, incluso aunque sea verdad que los dolores son idénticos en todos los humanos a sucesos neurofisiológicos humanos, no deseamos excluir la posibilidad de que, en alguna otra especie, pudiera ser que hubiera dolores que fueran idénticos a algún otro tipo de configuración neurofisiológica. En pocas palabras, parece excesivo esperar que cada *tipo* de estado mental sea idéntico a un *tipo* de estado neurofisiológico. Realmente, parece un tipo de «chauvinismo neuronal» (Block, 1978) suponer que sólo entidades con neuronas como nosotros pueden tener estados mentales.

Una tercera objeción «técnica» a la teoría de la identidad se deriva de la ley de Leibniz. Si dos sucesos son idénticos sólo si tienen todas sus propiedades en común, parece que los estados mentales no pueden ser idénticos a los estados físicos, porque los estados mentales tienen ciertas propiedades que los estados físicos no tienen (Smart, 1965; Shaffer, 1961). Por ejemplo, mi dolor está en mi pie, pero el estado neurofisiólogico correspondiente se extiende desde el pie hasta el tálamo e incluso más allá. Así que, ¿dónde está el dolor en realidad? Los teóricos de la identidad no tuvieron demasiados problemas con esta objeción. Señalaron que, en realidad, la unidad de análisis es la *experiencia* de tener dolor, y que, presumiblemente, esa experiencia (junto con la experiencia de la imagen de todo el cuerpo) tiene lugar en el sistema nervioso central (Smart, 1965). Sobre este extremo, me parece que los materialistas están completamente en lo cierto.

Una objeción técnica más radical a la teoría de la identidad la planteó Saul Kripke (1971), con el siguiente argumento modal: si fuera realmente una verdad que el dolor es idéntico a la estimulación de las fibras-C, debería ser una verdad necesaria, en el mismo sentido que el enunciado de identidad «El calor es idéntico al movimiento molecular» es una verdad necesaria. Esto es así porque, en ambos casos, las expresiones en cada lado del enunciado de identidad son «designadores rígidos». Con ello, lo que quiere decir es que cada expresión identifica el objeto al que se refiere en términos de sus propiedades esenciales. El sentimiento de dolor que tengo ahora es *esencialmente* un sentimiento de dolor porque

cualquier cosa idéntica a él debería ser dolor, y este estado cerebral es *esencialmente* un estado cerebral porque cualquier cosa idéntica a él habría de ser un estado cerebral. De modo que parece que el teórico de la identidad que pretende que los dolores son ciertos tipos de estados cerebrales, y que este dolor particular es idéntico a este estado cerebral particular, estaría obligado a mantener tanto que se trata de una verdad necesaria que, en general, los dolores son estados cerebrales, como que es una verdad necesaria que este dolor particular es un estado cerebral. Sin embargo, ninguna de estas afirmaciones parece correcta. No parece correcto afirmar ni que los dolores en general son necesariamente estados cerebrales, ni que mi dolor actual es necesariamente un estado cerebral, porque parece fácil imaginar que algún tipo de ser podría tener estados cerebrales como éstos sin tener dolores y tener dolores como éstos sin estar en este tipo de estados cerebrales. Es incluso posible concebir una situación en la que yo tuviera este mismo dolor sin tener este estado cerebral y en la que tuviera este estado cerebral sin tener dolor.

El debate sobre la fuerza de este argumento modal prosiguió durante algunos años y todavía continúa (Lycan, 1971, 1987; Sher, 1977). Desde el punto de vista de nuestro presente interés, deseo destacar el hecho de que se trata esencialmente de la objeción de sentido común bajo un ropaje sofisticado. La objeción de sentido común a cualquier teoría de la identidad es la de que no es posible identificar algo mental con algo no mental sin dejar a un lado lo mental. El argumento modal de Kripke es que la identificación de los estados mentales con los estados físicos habría de ser necesaria, y que, no obstante, no puede ser necesaria porque lo mental no podría ser necesariamente físico. Como dice Kripke, citando a Butler, «cada cosa es lo que es y no otra cosa».[6]

En cualquier caso, la idea de que cada tipo de estado mental es idéntico a algún tipo de estado neurofisiológico parecía realmente de-

6. No me interesa defender, en este capítulo, mi solución al problema mente-cuerpo, pero merece destacarse que no está sometida a esta objeción. Tanto Kripke como sus oponentes aceptan el vocabulario dualista, con su oposición entre «lo mental» y «lo físico», que yo rechazo. Una vez que esa contraposición se rechaza, mi punto de vista es el de que mi estado actual de dolor es un rasgo de nivel superior de mi dolor. Por lo tanto, es necesariamente idéntico a cierto rasgo de mi dolor, o sea, a él mismo. También necesariamente, no es idéntico a ningún otro de los rasgos de mi cerebro, aunque esté causado por ciertos sucesos de nivel inferior de mi cerebro. Es posible que tales rasgos pudieran ser causados por otros tipos de sucesos y podrían ser rasgos de otros tipos de sistemas. De modo que no hay ninguna conexión necesaria entre dolores y cerebros. Cada cosa es lo que es y no otra cosa.

masiado exigente. Pero pareció que la motivación filosófica subyacente al materialismo podría preservarse con una tesis mucho más débil, la tesis de que para cada instancia particular de un estado mental deberá haber algún suceso neurofisiológico particular al que esa instancia particular será idéntica. Tales puntos de vista se conocieron como «teorías de la identidad de las instancias» y reemplazaron pronto a las teorías de la identidad de tipos. De hecho, algunos autores pensaron que la teoría de la identidad de las instancias podría escapar a la fuerza de los argumentos modales de Kripke.[7]

IV. Teorías de la identidad de las instancias

Los teóricos de la identidad de las instancias heredaron la objeción de sentido común a las teorías de la identidad de tipos, la objeción de que todavía parecían preservar alguna forma de dualismo de propiedades. Además, también se enfrentaron a algunas dificultades idiosincrásicas.

Una de ellas fue la siguiente. Si dos personas que están en el mismo estado mental, están en estados neurofisiológicos diferentes, ¿qué es lo que hace de esos dos estados neurofisiológicos diferentes el mismo estado mental? Si usted y yo creemos que Denver es la capital de Colorado, ¿qué es lo que tenemos en común que convierte a nuestras distintas configuraciones neurofisiológicas en la misma creencia? Advirtamos que los teóricos de la identidad de las instancias no pueden dar la respuesta de sentido común a esta cuestión. No pueden decir que lo que hace que dos sucesos neurofisiológicos distintos pertenezcan al mismo tipo de suceso mental es el que posean el mismo tipo de rasgos mentales, porque el materialismo trata de conseguir, precisamente, la eliminación o reducción de esos rasgos mentales. Deben encontrar alguna respuesta no mentalista a la cuestión «¿Qué es lo que hay en dos estados neurofisiológicos diferentes que los hace instancias del mismo tipo de estado mental?». Dada toda la tradición en la que se movían, la única respuesta posible tenía que ser de estilo conductista. Su respuesta fue la de que un estado neurofisiológico era un estado mental particular en virtud de su función. Esto nos lleva de un modo natural al punto de vista que examinaremos a continuación.

7. Por ejemplo, McGinn (1977). McGinn defiende el argumento de Davidson en favor del «monismo anómalo», al que tanto él como Davidson consideran una versión de la teoría de la identidad de las instancias particulares.

V. El funcionalismo de la caja negra

Lo que convierte a dos estados neurofisiológicos en instancias del mismo tipo de estado mental es el hecho de que ambos cumplen la misma función en la vida global del organismo. La noción de función es algo vaga, pero los teóricos de la identidad de las instancias la especifican del siguiente modo. Dos instancias de diferentes estados cerebrales serían instancias del mismo tipo de estado mental si y sólo si los estados cerebrales tuvieran las mismas relaciones causales con los *inputs* estimulativos que recibe el organismo, con los otros estados «mentales», y con su *output* conductual (Lewis, 1972; Grice, 1975). Así, por ejemplo, mi creencia de que va a llover será un estado mío causado por la percepción de que se están formando muchas nubes y, junto con mi deseo de que no se cuele el agua a través de las ventanas, causará, a su vez, que yo las cierre. Adviértase que, al identificar los estados mentales en términos de sus relaciones causales —no sólo con los *inputs* estimulativos y los *outputs* conductuales, sino también con otros estados mentales—, los teóricos de la identidad de las instancias evitaron inmediatamente dos objeciones al conductismo. Una era la de que el conductismo había pasado por alto las relaciones causales de los estados mentales, y la otra era la de la circularidad, la de que las creencias tuvieran que ser analizadas en términos de deseos y los deseos en términos de creencias. El teórico de la identidad de las instancias, en su versión funcionalista, puede aceptar sin problemas esta circularidad, argumentando que la totalidad del sistema de conceptos puede atraparse en términos del sistema de relaciones causales.

El funcionalismo tiene un mecanismo técnico precioso con el que hacer completamente transparente ese sistema de relaciones sin invocar «entidades mentales misteriosas». Este mecanismo se denomina oración de Ramsey,[8] y funciona del modo siguiente: supongamos que Juan tiene la creencia de que *p*, y que es causada por su percepción de que *p*; y, junto a su deseo de que *q*, la creencia de que *p* causa su acción *a*. Dado que estamos definiendo las creencias en términos de sus relaciones causales, podemos eliminar el uso explícito de la palabra «creencia» en la oración anterior, y decir tan sólo que hay *algo* que está en tal y tal relación causal. Hablando formalmente, la manera de eliminar la mención explícita a la creencia es, simplemente, la de colocar una va-

8. Por el filósofo británico F. P. Ramsey (1903-1930).

riable «*x*», en lugar de cualquier expresión que se refiera a la creencia de Juan de que *p*, y colocar al comienzo de la oración un cuantificador existencial (Lewis, 1972). La historia completa de la creencia de Juan de que *p* puede ser contada, entonces, del siguiente modo:

($\exists$x) (Juan tiene *x* y *x* está causado por la percepción de que *p* y *x*, conjuntamente con un deseo de que *q*, causa la acción *a*)

Se supone que habrá oraciones de Ramsey adicionales que serán capaces de liberarse de los términos psicológicos que aún nos quedan, como «deseo» o «percepción». Una vez que las oraciones de Ramsey se detallan de este modo, sucede que el funcionalismo tiene la ventaja crucial de mostrar que no hay nada especialmente mental por lo que respecta a los estados mentales. Hablar de los estados mentales es, simplemente, hablar de un conjunto neutral de relaciones causales; y el «chauvinismo» aparente de las teorías de la identidad de tipos —es decir, el chauvinismo de suponer que sólo sistemas con cerebros como los nuestros son capaces de tener estados mentales— es evitado por este punto de vista mucho más «liberal».[9] Cualquier sistema, sin que importe de qué está constituido, podría tener estados mentales, con la única condición de que tuviera el tipo correcto de relaciones causales entre sus *inputs*, su funcionamiento interno y sus *outputs*. Este tipo de funcionalismo no dice nada acerca de cómo se las arregla la creencia para tener las relaciones causales que tiene. Se limita a tratar a la mente como una suerte de caja negra en la que suceden esos tipos de relaciones causales y, por esa razón, fue denominado en numerosas ocasiones «funcionalismo de la caja negra».

Las objeciones al funcionalismo de la caja negra revelaron la misma mezcla de sentido común y tecnicismos que hemos visto anteriormente. La objeción de sentido común fue la de que el funcionalista parece dejar a un lado el sentir cualitativo y subjetivo de, al menos, algunos de nuestros estados mentales. Involucradas en el ver un objeto rojo o en el tener un dolor de espalda, hay experiencias cualitativas muy específicas, y limitarse a describir esas experiencias en términos de sus relaciones causales deja a un lado esos *qualia* especiales. A este respecto, se ofreció la prueba siguiente: supongamos que un sector de la pobla-

9. La terminología de «chauvinismo» y «liberalismo» fue introducida por Ned Block (1978).

ción tuviera sus espectros de color invertidos de tal manera que, por ejemplo, la experiencia que ellos denominan «ver rojo» sería denominada «ver verde» por una persona normal; y lo que ellos denominan «ver verde» sería denominado «ver rojo» por una persona normal (Block y Fodor, 1972). Podríamos suponer que esta «inversión de espectro» es completamente indetectable por cualquiera de las pruebas convencionales del daltonismo, dado que el grupo anormal realiza exactamente las mismas discriminaciones de color en respuesta a exactamente los mismos estímulos que el resto de la población. Cuando se les pide que pongan en un montón los lápices rojos y en otro los lápices verdes, hacen exactamente lo mismo que haría el resto de nosotros; les parece *diferente* a ellos desde dentro, pero no hay modo alguno de detectar esa diferencia desde el exterior.

Ahora bien, si esta posibilidad es siquiera inteligible para nosotros —y seguramente lo es— el funcionalismo de la caja negra debe estar equivocado al suponer que las relaciones causales especificadas de un modo neutral son suficientes para explicar los fenómenos mentales; dado que tales especificaciones dejan a un lado un rasgo crucial de muchos fenómenos mentales, o sea, el cómo son sentidos.

Una objeción relacionada fue la de que una población enorme, por ejemplo, toda la población de China, podría comportarse de un modo tal que imitara la organización funcional de un cerebro humano hasta el punto de tener las relaciones *input-output* correctas y el patrón correcto de relaciones internas de causa y efecto. Aunque así fuera, el sistema todavía no sentiría nada como sistema. La totalidad de la población china no sentiría dolor sólo por imitar la organización funcional propia del dolor (Block, 1978).

Otra objeción aparentemente más técnica al funcionalismo de la caja negra iba dirigida al ingrediente de la «caja negra»: el funcionalismo así definido era incapaz de enunciar en términos materiales qué es lo que da a fenómenos materiales diferentes las mismas relaciones causales. ¿Cómo es posible que esas estructuras físicas diferentes sean causalmente equivalentes?

VI. LA INTELIGENCIA ARTIFICIAL FUERTE

En este punto, sucedió uno de los acontecimientos más excitantes en la totalidad de los dos mil años de historia del materialismo. La cien-

cia cognitiva en desarrollo proporcionó una respuesta a esta cuestión: las estructuras materiales diferentes pueden ser mentalmente equivalentes si son diferentes implementaciones de *hardware* del mismo programa de ordenador. En realidad, dada esta respuesta, podemos ver que la mente es simplemente un programa de ordenador y el cerebro uno de entre un rango indefinido de diferentes *hardwares* de ordenador (o *wetwares*) que puede poseer una mente. La mente es al cerebro como el programa es al *hardware* (Johnson-Laird, 1988). La inteligencia artificial y el funcionalismo se combinaron, y uno de los aspectos más sorprendentes de esa unión fue la consecuencia de que es posible ser un completo materialista sobre la mente y creer todavía, como Descartes, que el cerebro no importa nada para la mente. Dado que la mente es un programa de ordenador, y dado que un programa puede ser implementado en cualquier *hardware* (con la única condición de que el *hardware* sea lo suficientemente estable y poderoso como para ejecutar los pasos del programa), los aspectos específicamente mentales de la mente pueden ser especificados, estudiados y comprendidos sin saber cómo funciona el cerebro. Incluso aunque alguien sea un materialista, no tiene por qué estudiar el cerebro para estudiar la mente.

Esta idea dio origen a la nueva disciplina de la «ciencia cognitiva». Tendré algo más que decir sobre esto más adelante (en los capítulos 7, 9 y 10); en este momento sólo estoy trazando la historia reciente del materialismo. Tanto la disciplina de la inteligencia artificial como la teoría filosófica del funcionalismo convergieron en la idea de que la mente era sólo un programa de ordenador. He bautizado este punto de vista como «inteligencia artificial fuerte» (Searle, 1980a), y fue también denominado «funcionalismo del ordenador» (Dennett, 1978).

Las objeciones a la IA fuerte me parece que muestran la misma mezcla de objeciones de sentido común y objeciones más o menos técnicas que encontramos en otros casos. Las dificultades técnicas y las objeciones a la inteligencia articial, tanto en su versión débil como en su versión fuerte, son numerosas y complejas. No intentaré resumirlas aquí. En general, todas tienen que ver con ciertas dificultades a la hora de programar ordenadores de modo que pudieran satisfacer el *test* de Turing. Dentro de la IA misma, hubo siempre dificultades como el «problema del marco» (*the frame problem*) y la incapacidad para obtener análisis adecuados del «razonamiento no monotónico» que pudiera reflejar la conducta humana real. Desde fuera de la IA, se plantearon objeciones como las de Hubert Dreyfus (1972) que trataban de mostrar

que la manera en que la mente humana funciona es diferente de la manera en que funciona el ordenador.

La objeción de sentido común a la IA fuerte fue simplemente la de que el modelo computacional de la mente dejaba de lado cosas cruciales sobre la mente como la conciencia y la intencionalidad. Creo que el argumento más conocido contra la IA fuerte fue mi argumento de la habitación china (Searle, 1980a) que mostraba que un sistema podía instanciar un programa de modo que diera una simulación perfecta de alguna capacidad cognitiva humana, como la capacidad de comprender chino, aunque el sistema no comprendiera chino en absoluto. Imaginemos simplemente que alguien que no comprende chino en absoluto está encerrado en una habitación con multitud de símbolos chinos y un programa de ordenador para responder cuestiones en chino. El *input* del sistema consiste en símbolos chinos en forma de preguntas; el *output* del sistema consiste en símbolos chinos como respuesta a esas cuestiones. Podríamos suponer que el programa es tan bueno que las respuestas son indistinguibles de las de un hablante chino nativo. Aunque así sea, ni la persona que está dentro ni ninguna otra parte del sistema comprende literalmente chino, y, dado que el ordenador programado, *qua* ordenador, no tiene nada que no tenga este sistema, no entiende chino tampoco. Dado que un programa es algo puramente formal o sintáctico y dado que las mentes tienen contenidos mentales o semánticos, cualquier intento de producir una mente solamente con programas de ordenador deja a un lado los rasgos esenciales de la mente.

Además del conductismo, de las teorías de la identidad de tipos, de las teorías de la identidad de las instancias, el funcionalismo y la IA fuerte, hubo otras teorías en la filosofía de la mente dentro de la tradición materialista. Una de ellas, que se remonta a comienzos de los años sesenta en los trabajos de Paul Feyerabend (1963) y Richard Rorty (1965), ha sido recuperada recientemente en formas distintas por autores como P. M. Churchland (1981) y S. Stich (1983). Es el punto de vista de que los estados mentales no existen en absoluto, y es el denominado «materialismo eliminativo» al que ahora prestaré atención.

VII. MATERIALISMO ELIMINATIVO

En su forma más sofisticada, el materialismo eliminativo argumentaba del modo siguiente: nuestras creencias de sentido común sobre la

mente constituyen una suerte de teoría primitiva, una «psicología popular». Pero, como con cualquier teoría, las entidades postuladas por la teoría sólo pueden ser justificadas en la medida en que la teoría es verdadera. Del mismo modo en que el fracaso de la teoría de la combustión del flogisto eliminó cualquier justificación para creer en la existencia del flogisto, el fracaso de la psicología popular elimina la justificabilidad de las entidades de la psicología popular. De modo que, si la psicología popular fuera falsa, no estaríamos justificados para creer en la existencia de creencias, deseos, esperanzas, miedos, etc. De acuerdo con los materialistas eliminativos, es muy probable que la psicología popular resulte ser falsa. Parece probable que una «ciencia cognitiva» madura mostrará que nuestras creencias de sentido común sobre los estados mentales están completamente injustificadas. Este resultado tendrá la consecuencia de que las entidades que siempre hemos supuesto como existentes, nuestras entidades mentales ordinarias, no existen realmente. Y, por lo tanto, tenemos, en último término, una teoría de la mente que elimina la mente. De aquí la expresión «materialismo eliminativo».

Un argumento relacionado que se usa en favor del materialismo eliminativo me parece tan extraordinariamente malo que temo que lo debo haber entendido mal. En la medida en que puedo expresarlo, es esto lo que se argumenta:

Imaginemos que tuviéramos una ciencia neurobiológica perfecta. Imaginemos que tuviéramos una teoría que explicara realmente cómo funciona el cerebro. Tal ciencia cubriría el mismo campo que la psicología popular, pero sería mucho más poderosa. Además, parece altamente improbable que nuestros conceptos ordinarios de psicología popular, como creencia, deseo, esperanza, miedo, depresión, dolor, etc., casaran perfecta o remotamente con la taxonomía que nos proporcionara nuestra imaginaria ciencia neurobiológica perfecta. Con toda probabilidad, no quedaría lugar en esta neurobiología para expresiones como «creencia», «miedo», «esperanza» y «deseo», y no sería posible ninguna reducción adecuada de estos supuestos fenómenos.

Esta es la premisa. Aquí viene la conclusión:

Por lo tanto, las entidades supuestamente nombradas por las expresiones de la psicología popular, creencias, deseos, esperanzas, miedos, etc., no existen realmente.

Para ver hasta qué punto el anterior argumento es un mal argumento, imaginemos uno paralelo extraído de la física:

> Consideremos una ciencia existente, nuestra física teórica. Nos encontramos con una teoría que explica cómo funciona la realidad física, y que es enormemente superior a nuestras teorías de sentido común de acuerdo con todos los criterios habituales. La teoría física cubre el mismo campo que nuestras teorías de sentido común sobre clubs de golf, raquetas de tenis, vagones de tren, y casas de campo. Además, nuestros conceptos ordinarios de la física de sentido común como «club de golf», «raqueta de tenis», «ranchera marca Chevrolet» y «casa de campo» no casan, ni exacta ni remotamente, con la taxonomía de la física teórica. Simplemente, no hay uso en física teórica para ninguna de estas expresiones y no es posible ninguna reducción perfecta de tipos para estos fenómenos. La manera en que una física ideal —de hecho, la manera en la que nuestra física real— taxonomiza la realidad es realmente diferente de la manera en que la física ordinaria de sentido común taxonomiza la realidad.
>
> Por tanto, las casas de campo, las raquetas de tenis, los clubs de golf y las rancheras marca Chevrolet no existen en realidad.

No he visto publicada ninguna discusión de este error. Quizás es tan monumental que, simplemente, no se le presta ninguna atención. Descansa en la premisa, obviamente falsa, de que, para cada teoría empírica y su correspondiente taxonomía, a no ser que haya una reducción de tipo a tipo de las entidades taxonomizadas a las entidades de las mejores teorías de la ciencia básica, las entidades no existen. Si el lector tiene alguna duda sobre la falsedad de esta premisa, basta con que intente aplicarla a cualquiera de las cosas que ve a su alrededor —¡o a sí mismo![10]

De nuevo, nos encontramos con el mismo patrón de objeciones técnicas y de sentido común en el caso del materialismo eliminativo. Las objeciones técnicas tienen que ver con el hecho de que la psicología popular, si es una teoría, no es, sin embargo, un proyecto de investigación.

10. El argumento se encuentra en la obra de varios filósofos, por ejemplo, Steven Schiffer (1987) y Paul Churchland. Churchland da un sucinto enunciado de la premisa: «Si abandonamos la esperanza de la reducción, la eliminación emerge como la única alternativa posible» (1988).

En sí misma, no es un campo alternativo de investigación científica y, según los críticos de los materialistas eliminacionistas, es un hecho que éstos son, muchas veces, injustos en su ataque a la psicología popular. De acuerdo con sus defensores, la psicología popular no es tal tipo de mala teoría; es muy posible que muchos de sus supuestos centrales resulten ser verdaderos después de todo. La objeción de sentido común al materialismo eliminativo es, simplemente, la de que parece una locura. Parece una locura afirmar que yo nunca he sentido sed o deseos, que nunca he tenido dolor, que, en realidad, nunca he creído nada, o que mis creencias y mis deseos no juegan ningún papel en mi conducta. A diferencia de las teorías materialistas precedentes, el materialismo eliminativo no es que deje a un lado la existencia de la mente, sino que, desde el principio, niega la existencia de algo a lo que dejar a un lado. Cuando se enfrentan a la objeción de que el materialismo eliminativo parece demasiado absurdo como para merecer una consideración seria, sus defensores invocan casi invariablemente la maniobra de la época-heroica-de-la-ciencia (P. S. Churchland, 1987). Es decir, pretenden que abandonar la creencia de que tenemos creencias es análogo a abandonar la creencia en una Tierra plana o en las puestas del Sol, por ejemplo.

Vale la pena señalar, respecto a toda esta discusión, que en la historia del materialismo ha surgido cierta asimetría paradójica. Las primeras teorías de la identidad de tipos argumentaban que podíamos liberarnos de los misteriosos estados mentales cartesianos dado que tales estados no eran *nada más* que estados físicos (nada «además de» estados físicos); y lo argumentaban sobre el supuesto de que podría mostrarse que los tipos de estados mentales eran idénticos a los tipos de estados físicos, que nos encontraríamos con un ajuste entre los descubrimientos de la neurobiología y nociones ordinarias como las de dolor o creencia. Ahora bien, en el caso del materialismo eliminativo, es precisamente el supuesto fracaso de cualquier ajuste semejante lo que motiva la pretensión de eliminar esos estados mentales en favor de una neurobiología estricta. Los anteriores materialistas argumentaron que no había cosas tales como fenómenos mentales separados, porque los fenómenos mentales *son idénticos* a los estados cerebrales. Los materialistas más recientes argumentan que no hay cosas tales como los fenómenos mentales separados porque *no son idénticos* a los estados cerebrales. Me parece que este patrón es revelador, y lo que revela es el impulso a liberarse de los fenómenos mentales a toda costa.

VIII. La naturalización del contenido

Después de medio siglo de aparición recurrente de este patrón en los debates sobre el materialismo, podría suponerse que los materialistas y los dualistas pensarían que hay algo erróneo en los términos del debate. Pero, hasta ahora, parece que esta inducción no se le ha ocurrido a ninguna de las partes. Mientras escribo esto, el mismo patrón se repite en los intentos actuales de «naturalizar» el contenido intencional.

Estratégicamente, la idea es separar el problema de la conciencia del problema de la intencionalidad. Quizás se admita que la conciencia es irreductiblemente mental y, por tanto, que no está sujeta a tratamiento científico, pero es posible que, después de todo, la conciencia no importe mucho y que podamos arreglárnoslas sin ella. Sólo necesitamos naturalizar la intencionalidad, y «naturalizar la intencionalidad» quiere decir explicarla completamente en términos de —reducirla a— fenómenos no mentales, físicos. El funcionalismo fue uno de esos intentos de naturalizar el contenido intencional, y ha sido rejuvenecido por su unión con las teorías externalistas y causales de la referencia. La idea que está detrás de estos puntos de vista es la de que el contenido semántico, es decir, el significado, no puede estar completamente en el interior de nuestras cabezas, porque lo que hay en nuestras cabezas no basta para determinar cómo el lenguaje se relaciona con la realidad. Además de lo que haya en nuestras cabezas, «contenido estrecho», necesitamos un conjunto de relaciones causales reales con los objetos del mundo, necesitamos el «contenido amplio». Estos puntos de vista se desarrollaron, en primer lugar, en torno a problemas relacionados con la filosofía del lenguaje (Putnam, 1975b), pero es fácil ver cómo se extienden a los contenidos mentales en general. Si el significado de la oración «El agua es húmeda» no puede explicarse en términos de lo que sucede en el interior de las cabezas de los hablantes del castellano, entonces la creencia de que el agua es húmeda tampoco puede ser sólo asunto de lo que sucede en sus cabezas. En términos ideales, nos gustaría encontrar un análisis del contenido intencional que se estableciera tan sólo en términos de las relaciones causales entre la gente, por una parte, y los objetos y estados de cosas del mundo, por otra.

Un rival del intento externalista de naturalizar el contenido, y, en mi opinión, una explicación todavía menos plausible, es el punto de vista de que los contenidos intencionales pueden ser individualizados por su función teleológica, darwinista y biológica. Por ejemplo, mis de-

seos tendrán un contenido referido a agua o alimento si y sólo si funcionan para ayudarme a obtener agua o alimento (Millikan, 1984).

Hasta el momento, ningún intento de naturalizar el contenido ha producido una explicación (análisis, reducción) del contenido intencional que sea siquiera remotamente plausible. Consideremos el tipo más simple de creencia. Por ejemplo, creo que Flaubert fue mejor novelista que Balzac. ¿A qué debería parecerse un análisis de este contenido que se estableciera en términos de pura causalidad física, o selección natural darwinista, y que no utilizara términos mentales? No debe sorprender a nadie que estos análisis no puedan ni siquiera despegar del suelo.

De nuevo, tales concepciones naturalizadoras del contenido están sujetas tanto a objeciones de sentido común como a objeciones técnicas. El más famoso de los problemas técnicos es, probablemente, el problema de la disyunción (Fodor, 1987). Si cierto concepto es causado por cierto tipo de objeto, ¿cómo podemos dar cuenta de casos de identificación errónea? Si «caballo» es causado por caballos o por vacas que, erróneamente, son identificadas como caballos, ¿tenemos que decir que el análisis de «caballo» es disyuntivo, que significa o bien caballo, o bien cierto tipo de vacas?

Cuando escribo esto, los análisis naturalistas (externalistas, causales) del contenido están de moda. Terminarán fracasando por razones que, espero, ya resultan obvias. Acabarán dejando a un lado la subjetividad del contenido mental. Por medio de objeciones técnicas aparecerán contraejemplos, como los casos de disyunción, que deberán afrontarse usando ciertos trucos —predigo que serán relaciones nomológicas, o contrafácticos—, pero lo más que podemos esperar de los trucos, incluso si tienen éxito en bloquear los contraejemplos, será cierto paralelismo entre el resultado que ofrece el truco y las intuiciones sobre el contenido mental. Todavía estaremos lejos de la esencia del contenido mental.

No sé si alguien ha planteado ya la objeción obvia de sentido común al proyecto de naturalización del contenido mental, pero espero que esté claro, a partir de toda la discusión precedente, cuál será. Si nadie la ha planteado todavía, ahí va: cualquier intento de reducir la intencionalidad a algo no mental fracasará siempre porque deja a un lado la intencionalidad. Supongamos, por ejemplo, que tenemos un perfecto análisis causal y externalista de la creencia de que el agua es húmeda. Se proporciona este análisis estableciendo un conjunto de relaciones causales que un sistema mantiene con el agua y con la humedad y se especifican completamente esas relaciones sin ningún componente mental. El problema

es obvio: un sistema podría tener todas esas relaciones y, sin embargo, no creer que el agua es húmeda. Esto es sólo una extensión del argumento de la habitación china, pero apunta a una moraleja general: no es posible reducir el contenido intencional (o los dolores, o los *qualia*) a algo más, porque, si fuera posible, serían algo más, y no son algo más. El punto de vista opuesto al mío es enunciado de una manera muy escueta por Fodor: «Si el ser-sobre-algo (*aboutness*) es real, debe ser realmente algo más» (1987, p. 97). Por el contrario, el ser-sobre-algo (es decir, la intencionalidad) es real, y no es algo más.

Un síntoma de que hay algo radicalmente equivocado en el proyecto es que las nociones intencionales son inherentemente normativas. Establecen patrones de verdad, racionalidad, consistencia, etc., y no hay manera alguna en que esos patrones pudieran ser intrínsecos a un sistema que consistiera solamente en relaciones causales brutas, ciegas y no intencionales. No hay ningún componente normativo en la relación causal entre bolas de billar. Los intentos darwinistas y biológicos de naturalización del contenido tratan de evitar este problema apelando a lo que suponen que es el carácter inherentemente teleológico y normativo de la evolución biológica. Pero este es un error muy profundo. No hay nada normativo o teleológico en la evolución darwinista. De hecho, la mayor contribución de Darwin fue precisamente la eliminación del propósito y la teleología de la evolución, y su sustitución por formas puramente naturales de selección. La explicación de Darwin muestra que la teleología aparente de los procesos biológicos es una ilusión.

Una simple extensión de esta intuición consiste en señalar que nociones como las de «propósito» nunca son intrínsecas a los organismos biológicos (a menos, por supuesto, que esos organismos tengan procesos y estados intencionales y conscientes). E incluso nociones como «función biológica» son siempre relativas a un observador que asigna un valor normativo a los procesos causales. No hay ninguna diferencia *fáctica* entre decir:

1. El corazón causa el bombeo de la sangre

y decir:

2. La función del corazón es bombear la sangre.

Pero 2 asigna un estatus normativo a meros hechos causales brutos sobre el corazón, y lo hace por nuestro interés en la relación de este hecho con todo un montón de otros hechos, como nuestro interés en la supervivencia. En pocas palabras, los mecanismos darwinistas e, incluso, las mismas funciones biológicas están desprovistos por completo de propósito o teleología. Todos los rasgos teleológicos están, por entero, en la mente del observador.[11]

IX. La moraleja provisional

Mi propósito hasta ahora ha sido el de ilustrar un patrón recurrente en la historia del materialismo. Su representación gráfica se encuentra en el cuadro 2.1. No me he preocupado tanto de defender o refutar el materialismo como de examinar sus vicisitudes frente a ciertos hechos de sentido común sobre la mente, tales como que la mayor parte de nosotros estamos conscientes durante la mayor parte de nuestras vidas. Lo que encontramos en la historia del materialismo es una tensión recurrente entre el impulso a proporcionar una explicación de la realidad que deje a un lado cualquier referencia a los rasgos especiales de lo mental, tales como la conciencia y la subjetividad, y explicar, al mismo tiempo, nuestras intuiciones sobre la mente. Por supuesto, es imposible hacer ambas cosas. De modo que hay una serie de intentos, casi de carácter neurótico, para encubrir el hecho de que algún elemento crucial sobre los estados mentales está siendo dejado a un lado. Y cuando se señala que alguna verdad obvia está siendo negada por la filosofía materialista, los defensores de este punto de vista recurren casi invariablemente a ciertas estrategias retóricas diseñadas para mostrar que el materialismo debe ser correcto, y que el filósofo que ponga objeciones al materialismo debe apoyar alguna versión de dualismo, de misticismo, de esoterismo o de una tendencia general anticientífica. Pero la motivación inconsciente de todo esto, la motivación que, sin embargo, nunca llega a salir a la superficie, es el supuesto de que el materialismo es necesariamente inconsistente con la realidad y la eficacia causal de la conciencia, de la subjetividad, etc. Es decir, el supuesto básico que hay detrás del materialismo es esencialmente el supuesto cartesiano de que el materialismo implica antimentalismo y el mentalismo implica antimaterialismo.

11. En el capítulo 7 tendré algo más que decir sobre estos problemas.

CUADRO 2.1. *El patrón general mostrado por el materialismo reciente*

Teoría	Objeciones de sentido común	Objeciones técnicas
Conductismo lógico	No da cuenta de la mente: Objeciones del superactor/ superespartano	1. Circular. Requiere deseos para explicar creencias y viceversa 2. No puede especificar los condicionales 3. No da cuenta de la causalidad
Teoría de la identidad de tipos	No da cuenta de la mente o lleva al dualismo de propiedades	1. Chauvinismo neurológico 2. La ley de Leibniz 3. No puede explicar las propiedades mentales 4. Argumentos modales
Teoría de la identidad de las instancias	No da cuenta de la mente: *qualia* ausentes	No puede identificar los rasgos mentales de los contenidos mentales
Funcionalismo de la caja negra	No da cuenta de la mente: *qualia* ausentes y espectro invertido	No explica la relación de estructura y función
IA fuerte (Funcionalismo de máquina de Turing)	No da cuenta de la mente: la habitación china	El conocimiento humano es no representacional y, por ello, no computacional.
Materialismo eliminativo (Rechazo de la psicología popular)	Niega la existencia de la mente: injusto con la psicología popular	Defensa de la psicología popular
Naturalización de la intencionalidad	No da cuenta de la intencionalidad	Problema de la disyunción

Hay algo enormemente deprimente respecto a toda esta historia porque todo en ella parece innecesario y carente de propósito. Toda ella está basada en el supuesto falso de que el punto de vista sobre la realidad como algo completamente físico es inconsistente con el punto de vista de que el mundo contiene realmente estados mentales conscientes y subjetivos («cualitativos», «privados», «inmateriales», «no físicos») como los pensamientos y los dolores.

El aspecto más extraño de toda esta discusión es que el materialismo hereda el peor de los supuestos del dualismo. Al negar la pretensión dualista de que hay dos tipos de substancias en el mundo, o, al negar la pretensión del dualista de propiedades de que hay dos tipos de propiedades en el mundo, el materialismo acepta sin advertirlo las categorías y el vocabulario del dualismo. Acepta los términos en que Descartes plantea el debate. Acepta, en pocas palabras, la idea de que el vocabulario de lo mental y de lo físico, de lo material y de lo inmaterial, de la mente y el cuerpo está perfectamente bien tal y como está. Acepta la idea de que, si pensamos que existe la conciencia, estamos aceptando el dualismo. Lo que creo —como resulta obvio a partir de toda la discusión— es que el vocabulario, y las categorías concomitantes, son la fuente de nuestras dificultades filosóficas más profundas. En la medida en que usemos palabras como «materialismo», estamos obligados casi invariablemente a suponer que implican algo inconsistente con el mentalismo ingenuo. Lo que trato de mostrar es que, en este caso, es posible estar en misa y repicando. Uno puede ser un «materialista estricto» sin negar en modo alguno la existencia de fenómenos mentales (subjetivos, internos, intrínsecos, a menudo conscientes). Sin embargo, dado que mi uso de estos términos va contra más de trescientos años de tradición filosófica, probablemente sería mejor abandonar por completo este vocabulario.

Si uno tuviera que describir la motivación más profunda para abrazar el materialismo, podría decir simplemente que se trata tan sólo de un terror a la conciencia. Pero ¿debe ser así? ¿Por qué debe el materialista tener temor a la conciencia? ¿Por qué los materialistas no abrazan de buen grado la conciencia como una propiedad material más, junto a otras muchas? De hecho, algunos, como Armstrong y Dennett, pretenden hacer tal cosa. Pero lo hacen redefiniendo «conciencia» de tal modo que niegan el rasgo esencial de la conciencia, a saber: su carácter subjetivo. La razón más profunda para el miedo a la conciencia es que la conciencia tiene el rasgo esencialmente aterrador de la subjetividad. Los materialistas se muestran reluctantes a aceptar ese rasgo porque creen que aceptar la existencia de la conciencia subjetiva sería inconsistente con su concepción de cómo debe ser el mundo. Muchos creen que, dados los descubrimientos de las ciencias físicas, lo único que podemos aceptar es una concepción de la realidad que niegue la existencia de la subjetividad. De nuevo, como sucedía con «conciencia», una manera de enfrentarse a la situación es la de redefinir «subjetividad» de modo

que ya no signifique subjetividad, sino algo objetivo (como ejemplo, véase Lycan, 1990a).

Creo que todo esto equivale a un enorme error, y, en los capítulos 4, 5 y 6 examinaré en algún detalle el carácter y el estatus ontológico de la conciencia.

X. LOS ÍDOLOS DE LA TRIBU

He dicho anteriormente en este capítulo que explicaría por qué cierta cuestión que parecía muy natural era realmente incoherente. La cuestión es: ¿cómo es que fragmentos no inteligentes de materia producen inteligencia? En primer lugar, debemos tomar nota de la forma de la pregunta. ¿Por qué no nos planteamos la pregunta más tradicional: ¿cómo fragmentos inconscientes de materia producen conciencia? Esta pregunta me parece perfectamente coherente. Es una pregunta sobre cómo funciona el cerebro para causar estados mentales conscientes, incluso aunque las neuronas (o las sinapsis, o los receptores) individuales del cerebro no sean ellos mismos conscientes. Pero en la época actual somos reticentes a plantear la cuestión de esa manera porque carecemos de criterios «objetivos» de conciencia. La conciencia tiene una ontología subjetiva ineliminable, de modo que pensamos que es más científico replantear la cuestión como si fuera sobre la inteligencia, porque creemos que, respecto a la inteligencia, tenemos criterios impersonales y objetivos. Pero, en este punto, nos topamos inmediatamente con una dificultad. Si por «inteligencia» queremos decir algo que satisface los criterios objetivos de tercera persona para la inteligencia, la cuestión involucra una presuposición falsa. Porque si la inteligencia se define de un modo conductista no es el caso que las neuronas no sean inteligentes. Las neuronas, como casi todas las demás cosas del mundo, se comportan de acuerdo con ciertos patrones regulares y predecibles. Además, consideradas de cierta manera, las neuronas realizan un «procesamiento de información» extremadamente sofisticado. Reciben un conjunto rico de señales, provenientes de otras neuronas, en sus sinapsis dendríticas; procesan esa información en su cuerpo y envían la información a través de sus sinapsis axonales a otras neuronas. Si la inteligencia ha de definirse de un modo conductista, las neuronas son muy inteligentes bajo cualquier medida que se aplique. En pocas palabras, si nuestros criterios de inteligencia son completamente objetivos y de tercera persona —y todo el

interés en plantear la cuestión de este modo era el de obtener algo que satisficiera esas condiciones— la cuestión contiene una presuposición que es falsa en sus propios términos. La cuestión presupone, falsamente, que los fragmentos de materia no satisfacen los criterios de inteligencia.

No es sorprendente que la respuesta a la cuestión herede la misma ambigüedad. Hay dos conjuntos diferentes de criterios que se aplican a la expresión «conducta inteligente». Uno de estos conjuntos consiste en criterios «objetivos», o de tercera persona, que no son necesariamente de ningún interés en absoluto. Pero los otros conjuntos de criterios son esencialmente mentales e involucran el punto de vista de la primera persona. La «conducta inteligente» según el segundo conjunto de criterios involucra el pensar, y pensar es esencialmente un proceso mental. Ahora bien, si adoptamos los criterios de tercera persona para la conducta inteligente, entonces, por supuesto, los ordenadores —por no mencionar las calculadoras de bolsillo, coches, grúas, termostatos, y, en realidad, casi cualquier cosa del mundo— se comportan de un modo inteligente. Si somos consistentes al adoptar el test de Turing o algún otro criterio «objetivo» para la conducta inteligente, las respuestas a cuestiones como «¿Pueden fragmentos no inteligentes de materia producir conducta inteligente?», o incluso «¿Cómo lo hacen exactamente?» son ridículamente obvias. Cualquier termostato, calculadora de bolsillo o cascada de agua produce «conducta inteligente» y, en cada caso, sabemos cómo funciona. Ciertos artefactos se diseñan para comportarse como si fueran inteligentes y, dado que todo sigue las leyes de la naturaleza, todo tendrá alguna descripción bajo la que se comporta como si fuera inteligente. Pero este sentido de «conducta inteligente» no tiene ninguna relevancia psicológica en absoluto.

En pocas palabras, tendemos a escuchar tanto la pregunta como la respuesta como si oscilase entre dos polos diferentes: (a) ¿cómo fragmentos inconscientes de materia producen conciencia? (una pregunta perfectamente correcta que tiene una respuesta: en virtud de rasgos biológicos específicos —aunque en gran parte desconocidos— del cerebro); (b): ¿cómo fragmentos «no inteligentes» —¿de acuerdo con criterios de primera o de tercera persona?— de materia producen conducta «inteligente» —¿de acuerdo con criterios de primera o de tercera persona? Pero, en la medida en que utilicemos como criterios de inteligencia los criterios de tercera persona, la pregunta contiene una presuposición falsa, y esto es algo que se nos oculta porque tendemos a escuchar la pregunta bajo la interpretación (a).

APÉNDICE

¿EXISTE EL PROBLEMA DE LA PSICOLOGÍA POPULAR?

El objetivo del capítulo 2 no era tanto presentar mis propios puntos de vista como describir la historia contemporánea de una tradición filosófica. Quiero enunciar ahora algunos de mis puntos de vista sobre la llamada psicología popular (PP), puesto que no creo que hayan estado representados hasta ahora en la literatura sobre el tema. Las discusiones estándar, tanto a favor como en contra (Churchland, 1981; Stich, 1983; Horgan y Woodward, 1985, y Fodor, 1986) han estado dentro de la tradición.

Enunciaré por pasos el argumento como una serie de tesis y respuestas.

Tesis: la PP es una tesis empírica como cualquier otra, y como tal está sujeta a confirmación y disconfirmación empírica.

Respuesta: las capacidades efectivas que las personas tienen para habérselas consigo mismas y con los demás no tienen, en su mayor parte, forma proposicional. Son, en mi opinión, capacidades de Trasfondo. Cómo respondemos a las expresiones faciales, qué encontramos natural en la conducta, e incluso cómo entendemos las emisiones son, para poner un ejemplo, asuntos que tienen que ver, en gran medida, con el saber-cómo y no con teoría alguna. Se distorsionan esas capacidades si se piensa en ellas como en teorías. Véase el capítulo 8 para una elaboración de esta postura.

Tesis: sin embargo, se podrían enunciar correlatos o principios teóricos que subyacen en esas capacidades. Esto constituiría una psicología popular y sería con toda probabilidad falsa, puesto que, en general, las teorías populares son falsas.

Respuesta: Se puede, no sin alguna distorsión, enunciar un correlato teórico de una destreza práctica. Pero sería milagroso el que éstas sean, en general, falsas. Donde realmente son importantes, donde algo está en juego, las teorías populares tienen que ser, en general, verdaderas o, de lo contrario, no habrían sobrevivido. La física popular puede ser errónea sobre cuestiones periféricas como el movimiento de las esferas celestes o el origen de la Tierra, porque estas cosas no son demasiado importantes. Pero cuando se trata del modo en que se mueve nuestro cuerpo si uno salta desde un acantilado, o de lo que sucede si una gran roca le cae a uno encima, las teorías populares tienen que ser correctas o, de lo contrario, no habrían sobrevivido.

Tesis: se convierte ahora en un asunto específico de la ciencia cognitiva (CC) el decidir qué tesis de la PP son verdaderas y cuáles de sus compromisos ontológicos están justificados. La PP postula, por ejemplo, creencias y deseos para dar cuenta de la conducta, pero si resulta que la explicación de la conducta por parte de la CC es inconsistente con esto, entonces las creencias y deseos no existen.

Respuesta: casi todo el contenido de esta afirmación es erróneo. En primer lugar, no *postulamos* creencias y deseos para dar cuenta de nada. Simplemente experimentamos creencias y deseos conscientes. Piénsese en ejemplos de la vida real. Hace un día tórrido y vas conduciendo un camión a través del desierto en las afueras de Phoenix. No tienes aire acondicionado. No puedes recordar cuando has estado tan sediento y deseas con toda tu alma una cerveza fría. Ahora bien, ¿dónde está la «postulación» del deseo? Los deseos conscientes se experimentan. No se postulan en mayor medida que los dolores conscientes.

En segundo lugar, las creencias y los deseos causan algunas veces acciones, pero no hay una conexión esencial. La mayor parte de las creencias y deseos no dan como resultado acciones. Creo, por ejemplo, que el Sol está a 150 millones de kilómetros de distancia, y me gustaría tener cientos de miles de millones. ¿Cuáles de mis acciones explican esta creencia y este deseo? ¿Que si compro un billete para el Sol estaré

seguro de obtener un billete de 150 millones? ¿Que la próxima vez que alguien me dé 1.000 millones no los voy a rechazar?

Tesis: sea lo que sea, se postulen o no, es muy poco probable que haya una reducción adecuada de las entidades de la PP a la ciencia más básica de la neurobiología, de modo que parece que la eliminación es la única alternativa.

Respuesta: ya he dicho hasta qué punto este es un mal argumento. Muchos tipos de entidades reales desde los chalets adosados a las fiestas de cumpleaños, desde los tipos de interés a los partidos de fútbol, no soportan una reducción apropiada a entidades de alguna teoría fundamental. ¿Por qué deberían soportarla? Sospecho que tengo una «teoría» de las fiestas de cumpleaños —por lo menos en la medida en que tengo una teoría de la «psicología popular»— y las fiestas de cumpleaños consisten ciertamente en movimientos de moléculas; pero mi teoría de las fiestas de cumpleaños no es, ni por aproximación, tan buena como mi teoría de la física molecular, y no hay una reducción de las fiestas de cumpleaños a la taxonomía de la física. Pero a pesar de todo, las fiestas de cumpleaños existen realmente. La cuestión de la reductibilidad de tales entidades es irrelevante para la cuestión de su existencia.

¿Por qué habría de cometer alguien un error tan notable? Esto es: ¿por qué habría de suponer alguien que la «reducción apropiada» de creencias y deseos a la neurobiología es siquiera relevante para la existencia de creencias y deseos? La respuesta es que se está trazando una analogía falsa con la historia de ciertas partes de la física. Churchland piensa que «creencia» y «deseo» tienen el mismo estatus en la teoría de la psicología popular que tenían en la física «flogisto» y «fluido calórico». Pero la analogía se derrumba en toda suerte de modos: las creencias y los deseos, a diferencia del flogisto y el fluido calórico, no se postulaban como parte de una teoría especial, sino que se experimentan como parte de nuestra vida mental. Su existencia no es algo relativo a una teoría en mayor medida que lo es la existencia de chalets adosados, fiestas de cumpleaños, partidos de fútbol, tipos de interés, mesas o sillas. Siempre se pueden describir las creencias de sentido común que uno tiene sobre tales cosas como «teoría», pero la existencia de los fenómenos es anterior a la teoría. De nuevo, piénsese siempre en los casos efectivos. Mi teoría sobre las fiestas de cumpleaños incluiría cosas tales como que las grandes fiestas de cumpleaños tienden a ser más ruidosas que las

pequeñas, y mi teoría sobre los chalets adosados incluiría que tienden a extenderse más que otros tipos de casas. Tales «teorías» son, sin duda alguna, desesperadamente inadecuadas, y las entidades no soportan una reducción apropiada a la física, donde tenemos teorías mucho mejores que describen los mismos fenómenos. ¿Pero qué tiene todo esto que ver con la existencia de chalets adosados? Nada. De manera similar, la inadecuación de la psicología de sentido común y el fallo de la taxonomía de sentido común en encajar en la taxonomía de la ciencia del cerebro (esto es lo que se quiere decir cuando se habla de que no se ha logrado una «reducción apropiada») no tiene nada que ver con la existencia de creencias y deseos. Dicho brevemente: creencias y chalets adosados son totalmente distintos del flogisto porque su ontología no depende de la verdad de una teoría especial, y su irreductibilidad a una ciencia más fundamental es irrelevante para su existencia.

Tesis: sí, pero lo que estás diciendo pide la cuestión. Estás diciendo que creencias y deseos, al igual que las fiestas de cumpleaños y los chalets adosados, no son entidades teóricas —su base evidencial no se deriva de teoría alguna. ¿Pero no es este uno de los puntos en disputa?

Respuesta: pienso que es obvio que creencias y deseos se experimentan como tales y que, ciertamente, no se «postulan» para explicar la conducta porque no se postulan en absoluto. Sin embargo, ni siquiera las «entidades teóricas» alcanzan su legitimidad de la reductibilidad. Considérese la economía. Los tipos de interés, la demanda efectiva, la propensión marginal al consumo son todas ellas cosas a las que se hace referencia en economía matemática. Pero ninguno de los tipos de entidades en cuestión soporta una reducción apropiada a, por ejemplo, la física o la neurobiología. Y de nuevo, ¿por qué habrían de soportarla?

La reductibilidad es, en cualquier caso, una exigencia extraña para la ontología, puesto que clásicamente una manera de mostrar que una entidad *no* existe realmente ha sido reducirla a algo distinto. Así, las puestas de Sol se reducen a movimientos planetarios del sistema solar, lo cual mostraba que, como se concebía tradicionalmente, las puestas de Sol no existen. La apariencia de que el Sol se pone viene causada por algo distinto, esto es: por la rotación de la Tierra en relación con el Sol.

Tesis: con todo, es posible hacer una lista de afirmaciones de la psicología popular y ver que muchas de ellas son dudosas.

Respuesta: si se mira a las listas que se dan efectivamente, hay algo sospechoso en ellas. Si tuviera que hacer una lista de algunas proposiciones de la PP, incluiría cosas como las siguientes:

1. Las creencias pueden ser, en general, verdaderas o falsas.
2. Algunas veces las personas tienen hambre, y cuando tienen hambre desean a menudo comer algo.
3. Los dolores no son, a menudo, placenteros. Por esta razón, la gente trata a menudo de evitarlos.

Es difícil imaginar qué género de evidencia empírica podría refutar estas proposiciones. La razón es que, de acuerdo con una interpretación natural, no son hipótesis empíricas, o no son *sólo* hipótesis empíricas. Son más semejantes a los principios constitutivos de los fenómenos en cuestión. La proposición 1, por ejemplo, es más semejante a la «hipótesis» de que un *touchdown* en el fútbol norteamericano vale seis puntos. Si a uno se le dice que un estudio científico ha mostrado que los *touchdowns* valen efectivamente sólo 5,9999999999 puntos, nos daríamos cuenta enseguida que hay alguien aquí que está seriamente confundido. Es parte de nuestra definición ordinaria de *touchdown* el que vale seis puntos. Podemos cambiar la definición pero no podemos descubrir un hecho diferente. De forma similar, es parte de la definición de «creencia» el que las creencias son candidatos para la verdad y la falsedad. No podríamos «descubrir» que las creencias no son susceptibles de ser verdaderas o falsas.

Si se echa una ojeada a las listas de candidatos que se han dado para «leyes» de la PP, nos daremos cuenta de que tienden a ser o bien obviamente falsas o son principios constitutivos. Por ejemplo, Churchland (1981) incluye el principio de que «eliminando la confusión, la distracción, etc.», cualquiera que crea que p y si p entonces q, cree q (p. 209 en Lycan, 1990b). Como candidato para una creencia de sentido común esto es literalmente increíble. Si este principio fuese verdadero, entonces la demostración de teoremas no sería algo más difícil que el examen de las propias creencias (sin «confusión, distracción, etc.»). Es muy fácil refutar la PP si, para empezar, uno dice que consta de tales principios falsos.

Un candidato para principio constitutivo es el ejemplo de Churchland de que cualquiera que teme que p quiere que sea el caso que no p. ¿Cómo se buscaría evidencia empírica de que esto es falso? Es parte de

la definición de «temor». Así pues, el error profundo no es simplemente suponer que la PP es una teoría, sino que todas las proposiciones de la teoría son hipótesis empíricas.

Puesto que son constitutivas y no empíricas, la única manera de mostrar que son falsas sería mostrar que no tienen rango alguno de aplicación. Por ejemplo, los principios constitutivos de la brujería no se aplican a nada porque no hay brujas. Pero no se puede mostrar que los deseos conscientes y los dolores no existen de la manera en que se puede mostrar que las brujas no existen porque aquéllas son experiencias conscientes y no se puede hacer la distinción usual entre apariencia y realidad para las experiencias conscientes (sobre esta cuestión, véase el capítulo 3).

Se ha mostrado que una gran cantidad de creencias psicológicas de sentido común son falsas, y sin duda muchas más lo serán. Considérese un ejemplo espectacular: el sentido común nos dice que nuestros dolores están localizados en el espacio físico dentro de nuestros cuerpos, que, por ejemplo, un dolor en el pie está literalmente dentro del área del pie. Pero sabemos ahora que esto es falso. El cerebro forma una imagen corporal y los dolores, al igual que todas las sensaciones corporales, son parte de la imagen del cuerpo. El dolor-en-el-pie está literalmente en el espacio físico del cerebro.

Así pues, el sentido común estaba literalmente equivocado sobre algunos aspectos acerca de la localización de los dolores en el espacio físico. Pero incluso una falsedad extrema tal no muestra —y no puede mostrar— que los dolores no existen. Lo que es probable que suceda efectivamente, de hecho está sucediendo, es que el sentido común se complemente con conocimiento científico adicional. Por ejemplo, ahora reconocemos distinciones entre memoria a largo y a corto plazo, y entre éstas y la memoria icónica, y estas distinciones son el resultado de investigaciones neurobiológicas.

3. CÓMO ROMPER EL HECHIZO: CEREBROS DE SILICIO, ROBOTS CONSCIENTES Y OTRAS MENTES

La imagen del mundo como algo completamente objetivo ejerce una poderosa fascinación sobre nosotros, aunque es inconsistente con los hechos más obvios de nuestras experiencias. Dado que la imagen es falsa, tenemos que ser capaces de romper el hechizo. No conozco ninguna manera simple de hacerlo. Uno de los propósitos de este libro, sin embargo, es comenzar la tarea. En este capítulo, deseo describir algunos experimentos mentales que pondrán en cuestión la exactitud de la imagen. En un principio, el propósito de los experimentos mentales es el de poner en cuestión la concepción de lo mental como algo que tiene alguna conexión importante con la conducta.

Para comenzar a debilitar los fundamentos de todo este modo de pensar, deseo considerar algunas de las relaciones entre la conciencia, la conducta y el cerebro. La mayor parte de la discusión estará vinculada a los fenómenos mentales conscientes; pero el dejar al lado lo inconsciente en este momento no supone una gran limitación, porque, como argumentaré en detalle en el capítulo 7, no tenemos noción alguna de un estado mental inconsciente excepto en términos que se derivan de los estados conscientes. Para comenzar el argumento, emplearé un experimento de pensamiento que he mencionado en otro lugar (Searle, 1982). Este *Gedankenexperiment* es una vieja historia en filosofía, y no sé quién fue el primero en usarlo. Lo he empleado durante años en conferencias, y doy por sentado que cualquiera que piense sobre esos temas está destinado a que se le ocurran un día u otro ideas como estas.

I. CEREBROS DE SILICIO

Esta es la historia. Imagina que tu cerebro comienza a deteriorarse de tal forma que te vuelves ciego lentamente. Imagina que los desesperados médicos, ansiosos por aliviar tu estado, intentan, por todos los métodos posibles, que recuperes la visión. Como último recurso, intentan implantar pequeñas piezas de silicio en tu córtex visual. Imagina que, para tu sorpresa y la de ellos, las piezas de silicio restauran tu visión a su estado normal. Imagina ahora que tu cerebro, desgraciadamente, continúa deteriorándose y que los doctores continúan implantando más piececitas de silicio. El lector ya percibirá a dónde conduce este experimento mental: en último término, imaginamos que el cerebro ha sido reemplazado completamente por piezas de silicio; que, cuando agitas la cabeza, puedes oír el ruido que producen al chocar entre sí en el interior del cráneo. En una situación semejante, habría varias posibilidades. Una posibilidad lógica, que no debe excluirse por meras razones *a priori*, es, seguramente, la siguiente: tú continúas teniendo todas las clases de experiencias, pensamientos, recuerdos, etc., que tenías previamente, la secuencia de tu vida mental no se ve modificada. En este caso, estamos imaginando que las piezas de silicio no sólo tienen el poder de duplicar tus funciones *input-output*, sino también de duplicar los fenómenos mentales, conscientes o no, que son normalmente responsables de tus funciones *input-output*.

Me apresuro a añadir que no pienso, ni por un momento, que tal cosa sea ni siquiera remotamente empíricamente posible. Pienso que es empíricamente absurdo suponer que pudiéramos duplicar por completo los poderes causales de las neuronas en silicio. Pero esta es una afirmación empírica por mi parte. No es nada que pudiéramos establecer *a priori*. De modo que el experimento de pensamiento continúa siendo válido como un enunciado de posibilidad lógica o conceptual.

Imaginemos ahora algunas variaciones en el experimento de pensamiento. Una segunda posibilidad, que tampoco puede ser excluida sobre ninguna razón *a priori*, es la siguiente: a medida que el silicio se implanta progresivamente en tu cerebro menguante, descubres que el área de tu experiencia consciente se va reduciendo, sin que ello muestre ningún efecto en tu conducta externa. Descubres, para tu total sorpresa, que estás perdiendo realmente el control de tu conducta externa. Descubres, por ejemplo, que, cuando los médicos comprueban tu vista, les oyes decir «Estamos sosteniendo un objeto rojo frente a ti; por fa-

vor, dinos que ves». Deseas decir «No veo nada. Me estoy volviendo completamente ciego». Pero oyes que tu voz dice, de un modo que está absolutamente fuera de tu control, «Veo un objeto rojo frente a mí». Si conducimos este experimento mental hasta el límite, tenemos un resultado mucho más deprimente que la última vez. Imaginamos que tu experiencia consciente se reduce lentamente a nada, mientras tu conducta externa observable sigue siendo la misma.

Es importante, en estos experimentos de pensamiento, que siempre los concibamos desde la perspectiva de la primera persona. Preguntémonos: «¿Como qué me parecería a mí?» y veremos que es perfectamente concebible para cada uno de nosotros imaginar que nuestra conducta externa continúa siendo la misma, mientras que nuestros procesos internos de pensamiento consciente se reducen gradualmente a cero. Desde el exterior, parece a los observadores que todo va bien, pero, desde el interior, estamos muriendo poco a poco. En este caso, estamos imaginando una situación en la que, al final, estamos mentalmente muertos, en la que no tenemos ninguna vida mental consciente en absoluto, mientras que nuestra conducta externa observable continúa siendo la misma.

Es también importante en este experimento mental recordar nuestra estipulación de que nos estamos volviendo inconscientes mientras que nuestra conducta no se ve afectada. Para aquellos que sientan cierta perplejidad sobre cómo es posible que suceda, recordémosles simplemente esto: en la medida en que lo sabemos, la base de la conciencia está en ciertas regiones específicas del cerebro, tales como, quizás, la formación reticular. Y podemos suponer en este caso que estas regiones se están deteriorando gradualmente hasta el punto en que no hay conciencia en el sistema. Pero suponemos además que las piezas de silicio son capaces de duplicar las funciones *input-output* de la totalidad del sistema nervioso central, incluso cuando no queda conciencia en los restos del sistema.

Consideremos ahora una tercera variación. En este caso, imaginamos que la progresiva implantación de las piezas de silicio no produce ningún cambio en nuestra vida mental, sino que el sujeto es progresivamente más incapaz de poner en acción sus pensamientos, sentimientos e intenciones. En este caso, imaginamos que sus pensamientos, sentimientos, experiencias, recuerdos, etc., no se modifican, pero que la conducta externa observable se reduce hasta la parálisis total. Al final el sujeto sufre parálisis total incluso aunque su vida mental no sufra variación alguna. En este caso, oye a los médicos diciendo:

> Las piezas de silicio son capaces de mantener los latidos del corazón, la respiración y otros procesos vitales, pero es obvio que el paciente sufre muerte cerebral. Podríamos desconectar el sistema sin ningún problema, porque el paciente no tiene ninguna vida mental.

Ahora bien, en este caso, el paciente sabría que están totalmente equivocados. Esto es, querría gritar:

> ¡No, todavía estoy consciente! Percibo todo lo que sucede a mi alrededor. Sucede tan sólo que no puedo realizar ningún movimiento físico. He quedado completamente paralizado.

El objetivo de estas tres variedades de experimento de pensamiento es ilustrar las relaciones *causales* entre procesos cerebrales, procesos mentales y conducta externamente observable. En el primer caso, imaginábamos que las piezas de silicio tenían poderes causales equivalentes a los poderes del cerebro, e imaginábamos, por ello, que causaban tanto los estados mentales como la conducta que causan normalmente los procesos cerebrales. En el segundo caso, imaginábamos que la relación de mediación entre la mente y el patrón de conducta se rompía. En este caso, las piezas de silicio no duplicaban los poderes causales del cerebro para producir estados mentales conscientes, sólo duplicaban ciertas funciones *input-output* del cerebro. La vida mental subyacente se dejaba fuera.

En el tercer caso, imaginábamos una situación en la que el agente tenía la misma vida mental que antes, pero en la que los fenómenos mentales no tenían ninguna expresión conductual. De hecho, para imaginar este caso no necesitamos ni siquiera imaginar las piezas de silicio. Hubiera sido muy fácil imaginar una persona con los nervios motores seccionados de modo que estuviera completamente paralizada, mientras que la conciencia y otros estados mentales no se verían afectados. Algo como esto puede existir en la realidad clínica. Los pacientes que sufren el síndrome de Guillain-Barré están completamente paralizados, pero son también conscientes.

¿Cuál es la significación filosófica de estos tres experimentos de pensamiento? Me parece que hay un número de lecciones a aprender. La más importante es que nos proporcionan alguna ilustración sobre la relación entre la mente y la conducta. ¿Cuál es exactamente la importancia de la conducta para el concepto de mente? *Ontológicamente hablando, la conducta, el papel funcional y las relaciones causales son*

irrelevantes para la existencia de fenómenos mentales conscientes. Epistémicamente, captamos los estados mentales conscientes de los demás *parcialmente* a partir de su conducta. *Causalmente*, la conciencia sirve para mediar las relaciones causales entre estímulos de *input* y *output* de conducta y, desde un punto de vista *evolutivo*, la mente consciente funciona causalmente para controlar la conducta. Pero, hablando *ontológicamente*, los fenómenos en cuestión pueden existir completamente y tener todas sus propiedades esenciales independientemente de cualquier *output* conductual.

La mayoría de los filósofos que he estado criticando, aceptarían las dos proposiciones siguientes:

1. Los cerebros causan fenómenos mentales conscientes.
2. Hay alguna clase de conexión lógica o conceptual entre los fenómenos mentales conscientes y la conducta externa.

Pero lo que ilustran los experimentos mentales es que estas dos proposiciones no pueden mantenerse consistentemente con una tercera:

3. La capacidad del cerebro para causar la conciencia es conceptualmente diferente de su capacidad para causar la conducta motora. Un sistema podría tener conciencia sin conducta y conducta sin conciencia.

Pero, dada la verdad de 1 y de 3, debemos abandonar 2. De modo que el primer extremo que debemos deducir de nuestros experimentos de pensamiento es lo que podríamos denominar «el principio de la independencia de conciencia y conducta». En el caso número 2, imaginábamos la circunstancia en la que la conducta no se veía afectada, aunque desaparecían los estados mentales, de modo que la conducta no es una condición suficiente para los fenómenos mentales. En el caso número 3, imaginábamos la circunstancia en la que los fenómenos mentales estaban presentes, pero en la que desaparecía la conducta, de modo que la conducta tampoco es una condición necesaria para la presencia de lo mental.

Los experimentos de pensamiento ilustran dos extremos más. En primer lugar, la ontología de lo mental es esencialmente una ontología de la primera persona. No es más que una manera colorista de decir que todo estado mental ha de ser el estado mental de *alguien*. Los estados

mentales sólo existen como fenómenos subjetivos, de primera persona. Y el otro extremo relacionado es que, hablando epistemológicamente, el punto de vista de la primera persona es completamente diferente del punto de vista de la tercera. Es fácil imaginar casos, como los que ilustran nuestros experimentos mentales, en los que, desde un punto de vista de la tercera persona, alguien podría no ser capaz de decir si yo tengo estados mentales en absoluto. Podría incluso pensar que yo estaba inconsciente y, con todo, podría ser el caso que yo estuviera completamente consciente. Desde el punto de vista de la primera persona, no es posible cuestionar que estoy consciente, incluso aunque no hubiera pruebas accesibles para la tercera persona.

II. Robots conscientes

Deseo introducir un segundo experimento de pensamiento para apoyar las conclusiones proporcionadas por el primero. El propósito de este experimento de pensamiento, como en el primer caso, es usar nuestras intuiciones para tratar de trazar una línea de separación entre los estados mentales y la conducta. Imaginemos que estamos diseñando robots para que trabajen en una cadena de montaje. Imaginemos que nuestros robots son realmente demasiado elementales y tienden a estropearlo todo cuando se enfrentan a los aspectos más complejos de su tarea. Pero imaginemos que sabemos lo bastante sobre los rasgos electroquímicos de la conciencia humana para saber cómo producir robots que posean un nivel más bien bajo de conciencia, de modo que podemos diseñar y fabricar robots conscientes. Imaginemos, además, que estos robots conscientes son capaces de realizar discriminaciones que los robots inconscientes no podrían hacer, de modo que realizan una tarea mejor en la cadena de montaje. ¿Hay algo de incoherente en la historia precedente? He de decir que, de acuerdo con mis «intuiciones», es perfectamente coherente. Por supuesto, se trata de ciencia ficción, pero, entonces, muchos de los experimentos de pensamiento más importantes en filosofía y ciencia son, precisamente, ciencia ficción.

Imaginemos ahora un rasgo desafortunado de nuestros robots conscientes: supongamos que son absolutamente desgraciados. Podemos suponer, de nuevo, que nuestra neurofisiología es suficiente para que establezcamos que se sienten muy mal. Imaginemos ahora que le damos a nuestro grupo de investigación en robótica la siguiente tarea: di-

señar un robot que tenga la capacidad de hacer las mismas discriminaciones que los robots conscientes, por más que sean completamente inconscientes. Podremos, entonces, permitir que los robots desgraciados se retiren a una vejez más satisfactoria y placentera. Me parece que se trata de un proyecto de investigación bien definido; y podemos suponer que, operacionalmente hablando, nuestros científicos tratan de diseñar un robot con un *hardware* que saben que no causará o mantendrá la conciencia, aunque tendrá las mismas funciones *input-output* que el robot que tiene un *hardware* que causa y mantiene la conciencia. Así pues, podríamos suponer que tienen éxito, que fabrican un robot que es completamente inconsciente, pero que tiene los poderes y las capacidades conductuales que son absolutamente idénticas a las del robot consciente.

El propósito del experimento, como los anteriores, es mostrar que, en la medida en que está implicada la ontología de la conciencia, la conducta es simplemente irrelevante. Podríamos tener *conducta idéntica* en dos diferentes sistemas, uno de los cuales es consciente y el otro totalmente inconsciente.

III. El EMPIRISMO Y «EL PROBLEMA DE LAS OTRAS MENTES»

Muchos filósofos de mentalidad empirista sentirán cierta desazón por estos dos experimentos de pensamiento, especialmente por el primero. Les parecerá que estoy apelando a la existencia de hechos empíricos sobre los estados mentales de un sistema que no son decidibles por ningún medio empírico. Su concepción de los medios empíricos para decidir la existencia de hechos mentales reside enteramente en el presupuesto de la evidencia conductual. Creen que la única evidencia que tenemos para atribuir estados mentales a otros sistemas es la conducta de esos sistemas.

En esta sección, deseo continuar la exposición del problema de las otras mentes que comenzó en el capítulo 1. Parte de mi propósito será mostrar que no hay nada de incoherente u objetable en las implicaciones epistemológicas de los dos experimentos de pensamiento que acabo de describir, pero mi propósito principal será proporcionar un análisis de la base «empírica» para suponer que las otras personas y los animales superiores tienen fenómenos mentales más o menos parecidos a los de cada uno de nosotros.

Vale la pena poner de relieve al principio de la discusión que en la historia de la filosofía empirista y de la filosofía de la mente hay una ambigüedad sistemática en el uso de la palabra «empírico», una ambigüedad entre un sentido ontológico y un sentido epistemológico. Cuando la gente habla de hechos empíricos, muchas veces quieren decir hechos reales y contingentes del mundo en tanto que opuestos, por ejemplo, a los hechos de las matemáticas o hechos de la lógica. Pero, a veces, cuando la gente habla de hechos empíricos, quiere decir hechos que son comprobables por medios propios de la tercera persona, es decir, por «medios empíricos» y «métodos empíricos». Ahora bien, esa ambigüedad sistemática en el uso de la pálabra «empírico» sugiere algo que es ciertamente falso: que todos los hechos empíricos, en el sentido ontológico de ser hechos del mundo, son igualmente accesibles, desde el punto de vista epistemológico, a todos los observadores competentes. Sabemos independientemente que esto es falso. Hay muchos hechos empíricos que no son igualmente accesibles a todos los observadores. Las secciones previas nos ofrecieron algunos experimentos mentales diseñados para mostrar esto, pero tenemos de hecho algunos datos empíricos que sugieren exactamente el mismo resultado.

Consideremos el siguiente ejemplo.[1] Podemos imaginar, con alguna dificultad, cómo sería ser un pájaro que vuela. Digo «con alguna dificultad» porque, por supuesto, la tentación es siempre la de imaginar cómo sería si nosotros voláramos y no, hablando estrictamente, cómo sería volar *para un pájaro*. Pero ahora alguna investigación reciente nos dice que ciertos pájaros se desplazan detectando el campo magnético de la Tierra. Supongamos que, del mismo modo que el pájaro tiene una experiencia consciente de aleteo, o de sentir la presión del viento sobre su cabeza y su cuerpo, también tiene la experiencia consciente de sentir la corriente de magnetismo a través de su cuerpo. ¿Cómo es sentir el magnetismo a través del cuerpo? En este caso, no tengo la más remota idea de cómo siente un pájaro o, daría lo mismo, un ser humano, una corriente magnética procedente del campo magnético de la Tierra. Es, así lo considero, un hecho empírico si los pájaros que se desplazan detectando el campo magnético tienen de hecho una experiencia consciente de la detección del campo magnético. Pero el carác-

1. Siguiendo el espíritu del artículo de Thomas Nagel «What Is It Like to Be a Bat?» (1974).

ter cualitativo exacto de este hecho empírico no es accesible a las formas habituales de una prueba empírica. En realidad, ¿por qué deberían serlo? ¿Por qué debemos suponer que todos los hechos del mundo son igualmente accesibles a las pruebas convencionales y objetivas de tercera persona? Si pensamos sobre ello, el supuesto es obviamente falso.

He dicho que este resultado no es tan deprimente como podría parecer. La razón es simple. Aunque en algunos casos no tenemos un acceso idéntico a ciertos hechos empíricos por su subjetividad intrínseca, en general tenemos métodos indirectos para llegar a los mismos hechos empíricos. Consideremos el siguiente ejemplo. Estoy completamente convencido de que mi perro, al igual que otros animales superiores, tiene estados mentales conscientes, como experiencias visuales, sentimientos de dolor y sensaciones de sed y hambre, de frío y de calor. ¿Por qué estoy tan convencido de ello? La respuesta convencional apela a la conducta del perro, porque al observar esa conducta infiero que tiene estados mentales como los míos. Creo que esta respuesta es un error. No es sólo porque el perro se comporta del modo que corresponde a tener estados mentales conscientes, sino también porque puedo ver que la base causal de la conducta en la fisiología del perro es relevantemente similar a la mía propia. No se trata sólo de que el perro tiene una estructura como la mía y que tiene conducta que es interpretable de modos análogos a la manera en que interpreto la mía. Más bien, es la combinación de esos dos hechos, del hecho de que puedo ver que su conducta es la apropiada y de que tiene la *relación causal* apropiada en la fisiología subyacente. Puedo ver, por ejemplo, que éstos son los oídos del perro, ésta su piel, éstos sus ojos; que si pinchamos su piel, nos encontramos con la conducta correspondiente a pinchar la piel; si gritamos en sus oídos, nos encontramos con la conducta correspondiente a gritar en los oídos.

Es importante poner de relieve que no necesitamos tener una teoría anatómica y fisiológica elaborada y sofisticada de la estructura del perro, sino, por así decirlo, una simple anatomía y fisiología «popular» —la capacidad de reconocer la estructura de la piel, ojos, dientes, pelo, nariz, etc., y la capacidad de suponer que el papel causal que juegan en sus experiencias es relevantemente similar al papel que tales rasgos juegan en las experiencias propias. En realidad, incluso la descripción de ciertas estructuras como «ojos» u «oídos» ya implica que les estamos atribuyendo funciones y poderes causales similares a nuestros pro-

pios ojos y oídos. En pocas palabras, aunque yo no tengo acceso directo a la conciencia del perro, sin embargo sí que me parece que es un hecho empírico bien comprobado que los perros tienen conciencia, y está comprobado por evidencia empírica que es abrumadora. No tengo el mismo grado de confianza en relación a animales muy inferiores en la escala filogenética. No tengo ni idea sobre si las moscas, los cangrejos o los gusanos tienen conciencia. Me parece que es razonable dejar esas cuestiones en manos de los neurofisiólogos. Pero ¿qué tipo de evidencia buscarían los neurofisiólogos? He aquí otro experimento de pensamiento que podríamos imaginar.

Supongamos que tuviéramos una explicación de las bases neurofisiológicas de la conciencia en los seres humanos. Supongamos que tuviéramos causas precisas, aislables neurofisiológicamente, de la conciencia en los seres humanos, tales como que la presencia de los fenómenos neurofisiólogicos relevantes fuera una condición necesaria y suficiente para la conciencia. Quien los tiene, adquiere conciencia, quien los pierde, pierde la conciencia. Supongamos ahora que algunos animales tienen este fenómeno, denominémoslo «*x*» para abreviar, y que otros carecen de él. Supongamos que se hubiera descubierto que *x* ocurría en todos aquellos animales, como nosotros mismos, los monos, los perros, etc., respecto de los que estamos convencidos que tienen conciencia sobre la base de su fisiología más obvia, y que *x* estaba totalmente ausente en animales, como las amebas, respecto de las que no sentimos ninguna inclinación a atribuirles conciencia. Supongamos, además, que la eliminación de *x* de la neurofisiología de un ser humano produjera inmediatamente un estado de inconsciencia, y que su reintroducción produjera conciencia. En tal caso, me parece que podríamos suponer razonablemente que la presencia de *x* jugaba un papel crucial en la producción de la conciencia, y que este descubrimiento nos capacitaría para decidir casos dudosos respecto a si ciertos animales tienen o carecen de conciencia. Si las serpientes tuvieran *x*, y ciertos arácnidos carecieran de él, podríamos inferir razonablemente que esos arácnidos se estaban comportando siguiendo simples tropismos, mientras que las serpientes tenían conciencia en el mismo sentido en el que nosotros, los perros y los monos la tenemos.

Ni por un momento supongo que la neurofisiología de la conciencia ha de ser tan simple. Me parece mucho más probable que nos encontraremos con una enorme variedad de formas de neurofisiologías de la conciencia y que, en cualquier situación experimental real, buscaría-

mos evidencia independiente de la existencia de mecanismos similares a los tropismos para explicar la conducta aparentemente dirigida a fines de los organismos que carecieran de conciencia. Lo que pretende el ejemplo es sólo mostrar que podemos tener medios indirectos, de una clase objetiva, empírica, de tercera persona, para captar fenómenos empíricos que son intrínsecamente subjetivos y, por tanto, inaccesibles a pruebas directas de tercera persona.

No debe pensarse, sin embargo, que hay algo de segunda categoría o imperfecto en los métodos de tercera persona para descubrir estos hechos empíricos, subjetivos y de primera persona. Los métodos descansan en un simple principio que usamos en la vida ordinaria: *las mismas causas, los mismos efectos* y *causas similares, efectos similares.* Podemos ver inmediatamente, en el caso de otros seres humanos, que las bases causales de sus experiencias son virtualmente idénticas a las bases causales de nuestras experiencias. Es por ello por lo que, en la vida ordinaria, no hay ningún problema de «las otras mentes». Los animales proporcionan una buena prueba para este principio porque, por supuesto, no son fisiológicamente idénticos a nosotros, aunque sean similares en ciertos aspectos importantes. Tienen boca, oídos, ojos, nariz, etc. Por esta razón no dudamos realmente de que tengan las experiencias que van unidas a estos tipos diversos de órganos. Hasta ahora, todas estas consideraciones son precientíficas. Pero supongamos que pudiéramos identificar causas exactas de la conciencia para el caso de los seres humanos, y que, luego, pudiéramos descubrir exactamente las mismas causas en otros animales. Si fuera de ese modo, me parece que habríamos establecido concluyentemente que otras especies tienen exactamente el mismo tipo de conciencia que tenemos nosotros, porque podemos presumir que las mismas causas producen los mismos efectos. No sería una especulación arbitraria, porque tendríamos una muy buena razón para suponer que esas causas producirían los mismos efectos en otras especies.

En la práctica real, los libros de texto de neurofisiología proporcionan continuamente información, por ejemplo, sobre cómo la percepción de los colores por parte del gato es semejante a, y diferente de, la humana *e incluso a la de otros animales.* ¡Qué irresponsabilidad tan impresionante! ¿Cómo pueden sus autores pretender haber solucionado tan fácilmente el problema de las otras mentes para el gato? La respuesta es que el problema se soluciona para la visión del gato, una vez que sabemos exactamente cómo los órganos visuales

del gato son similares a, y diferentes de, los nuestros y a los de otras especies.[2]

Una vez que comprendamos las bases causales de la adscripción de estados mentales a otros animales, varios problemas tradicionales del escepticismo sobre las «otras mentes» tienen una fácil solución. Consideremos el famoso problema del espectro invertido que mencioné en el capítulo 2. Se dice a menudo que, por todo lo que sabemos, un sector de la población podría tener una inversión verde/rojo de tal modo que, aunque hicieran las mismas discriminaciones conductuales que el resto de nosotros, las experiencias reales que tendrían cuando vieran verde, a las que denominarían «de ver verde», serían experiencias que nosotros denominaríamos, si las tuviéramos, «de ver rojo», y viceversa. Supongamos que hubiéramos descubierto de hecho que un sector de la población tiene realmente invertidos los receptores de verde y de rojo, y conectados de la forma apropiada al resto de sus órganos visuales, de tal forma que tenemos evidencia neurofisiológica abrumadora de que, aunque sus discriminaciones son las mismas que las nuestras, realmente tienen diferentes experiencias subyaciendo en ellas. Esto no constituiría un problema de escepticismo filosófico, sino una hipótesis neurofisiológica bien determinada. Pero, entonces, si no existe ese tipo de sector en la población, si todas las personas que no son ciegas a los colores tienen las mismas rutas perceptivas para el verde y el rojo, tenemos una sólida evidencia empírica de que las cosas les parecen a las demás personas como nos parecen a nosotros. Una nube de escepticismo filosófico está condensada en una gota de neurociencia.

Adviértase que esta solución al «problema de las otras mentes», una solución que usamos en ciencia y en la vida ordinaria, nos proporciona condiciones suficientes, aunque no necesarias, para la adscripción correcta de fenómenos mentales a otros seres. Necesitaríamos, como sugería antes en este capítulo, una teoría neurobiológica de la conciencia mucho más rica que nada que podamos imaginar para suponer que podríamos aislar condiciones necesarias de la conciencia. Tengo absoluta confianza en que la mesa frente a mí, el ordenador que uso todos

2. Por ejemplo: «Como podría esperarse, han sido detectadas en diversos animales, que incluyen el mono, la ardilla de tierra, y algunos peces, células cuyos campos receptores están específicamente codificados para el color. *Estos animales, en oposición al gato, poseen una excelente visión de los colores*, y un intrincado mecanismo neurológico para el procesamiento del color» (Kuffler y Nicholls, 1976, p. 25, las cursivas son mías).

los días, el bolígrafo con el que escribo, el magnetófono frente al que hablo, carecen de conciencia. Pero, por supuesto, ni yo ni nadie más puede *probar* que carecen de conciencia.

IV. Resumen

En este capítulo, he tenido dos objetivos hasta ahora: en primer lugar, he intentado argumentar que la conducta es simplemente irrelevante, en la medida en que está implicada la ontología de la mente. Por supuesto, en la vida real nuestra conducta es crucial para nuestra misma existencia, pero cuando estamos examinando la existencia de nuestros estados mentales como estados mentales, la conducta correspondiente no es ni necesaria ni suficiente para su existencia. En segundo lugar, he intentado comenzar a romper el encanto de trescientos años de discusiones epistemológicas sobre el «problema de las otras mentes», de acuerdo con las cuales la conducta es la única base que tenemos para conocer la existencia de otras mentes. Esto me parece obviamente falso. La conducta es relevante para el descubrimiento de los estados mentales de los demás tan sólo por la *conexión* entre la conducta y la estructura causal de otros organismos.

Una última observación es igualmente importante: excepto cuando hacemos filosofía, no hay realmente ningún «problema» sobre las otras mentes, porque no mantenemos una «hipótesis», «creencia» o «supuesto» de que las otras personas tienen conciencia, y que las sillas, las mesas, los ordenadores y los coches no tienen conciencia. Más bien, tenemos cierto Trasfondo de formas de conducta, ciertas capacidades de Trasfondo, que son constitutivas de nuestras relaciones con la conciencia de otras personas. Es típico de la filosofía que los problemas escépticos surgen a veces cuando los elementos del Trasfondo son tratados como hipótesis que hubieran de ser justificadas. No mantengo la «hipótesis» de que mi perro o el jefe de mi departamento tienen conciencia, y, por lo tanto, la cuestión no surge más que en el debate filosófico.

V. Intencionalidad intrínseca, como-si y derivada

Antes de seguir más adelante, necesito introducir algunas distinciones simples que han estado implícitas en lo que he dicho hasta ahora,

pero que han de hacerse explícitas para lo que viene a continuación. Para introducirlas, consideremos las semejanzas y diferencias entre las diversas clases de condiciones de verdad de las oraciones que usamos para atribuir fenómenos mentales intencionales. Consideremos las semejanzas y diferencias entre las siguientes:

1. Estoy sediento ahora, realmente sediento, porque no he bebido nada durante todo el día.
2. Mi césped está sediento, realmente sediento, porque no ha sido regado en una semana.
3. En inglés, «I am very thirsty» quiere decir «Estoy muy sediento».

La primera de estas oraciones es usada literalmente para adscribir un estado mental, intencional y real a uno mismo. Si emito esta oración, realizando un enunciado verdadero, existe en mí una sensación consciente de sed que convierte el enunciado en verdadero. Esta sensación tiene intencionalidad porque involucra un deseo de beber. Pero la segunda oración es completamente diferente. La oración 2 se usa sólo metafórica o figuradamente para adscribir sed a mi césped. Mi césped, al carecer de agua, está en una situación en la que yo estaría sediento, de modo que, figuradamente, lo describo *como si* estuviera sediento. Por analogía, puedo afirmar inocuamente que el césped está sediento incluso si no supongo ni por un momento que el césped está literalmente sediento. La tercera oración es semejante a la primera en que adscribe literalmente intencionalidad, pero es semejante a la segunda y desemejante de la primera en que la intencionalidad descrita no es intrínseca al sistema.

El primer tipo de adscripción adscribe intencionalidad *intrínseca.* Si tal enunciado es verdadero, debe existir realmente un *estado intencional en el objeto de la adscripción.* La segunda oración no adscribe ninguna intencionalidad en absoluto, ni intrínseca ni de otra forma; se usa sólo para hablar metafórica o figuradamente. Por lo tanto, diré que la «intencionalidad» en la adscripción es meramente *como-si* y no intrínseca. Para evitar confusiones, es importante poner de relieve que la intencionalidad *como-si* no es un tipo de intencionalidad, más bien un sistema que tiene intencionalidad *como-si* es como-si-tuviera-intencionalidad. En el tercer caso, adscribo literalmente intencionalidad a la oración inglesa, es decir, la oración inglesa dice literalmente lo que digo que significa. Pero la intencionalidad de la oración inglesa no es

intrínseca a la oración particular concebida sólo como un objeto sintáctico. Esa misma secuencia podría haber significado algo muy diferente o nada en absoluto. *Los hablantes* del inglés la pueden utilizar para expresar la intencionalidad *de ellos*. El significado lingüístico es una forma de intencionalidad, pero no se trata de intencionalidad intrínseca. Se deriva de la intencionalidad intrínseca de los usuarios del lenguaje.

Podemos resumir estas observaciones de la manera siguiente: la intencionalidad intrínseca es un fenómeno que los seres humanos y algunos otros animales tienen como parte de su naturaleza biológica. No es asunto de cómo son usados o cómo se conciben, o deciden describirse, a sí mismos. Se trata tan sólo de un hecho bruto sobre tales criaturas que, por ejemplo, a veces se sienten *hambrientas* o *sedientas*, que *ven* cosas, que *temen* cosas, etc. Todas las expresiones en cursiva de la oración anterior se usan para hacer referencia a estados intencionales intrínsecos. Es muy conveniente usar el vocabulario de la intencionalidad para hablar de sistemas que no la poseen, pero que se comportan como si la tuvieran. Digo que mi termostato *percibe* los cambios de temperatura; digo que mi carburador *sabe* cuándo enriquecer la mezcla; digo de mi ordenador que tiene una *memoria* mayor que la del ordenador que tuve el año pasado. Todas estas atribuciones son completamente inofensivas y, sin ninguna duda, acabarán por producir nuevos significados literales a medida que se vaya perdiendo el contenido metáforico. Pero es importante subrayar que esas atribuciones son psicológicamente irrelevantes, dado que no implican la presencia de ningún fenómeno mental. La intencionalidad descrita en todos estos casos es puramente *como-si*.

Los casos del tercer tipo se convierten en interesantes por el hecho de que, a menudo, adjudicamos literalmente propiedades intencionales a los fenómenos no mentales. No hay nada metafórico o *como-si* en la afirmación de que ciertas oraciones *significan* ciertas cosas, o que ciertos mapas son *representaciones* correctas del estado de California, o que ciertos retratos son *retratos de* Winston Churchill. Estas formas de intencionalidad son reales, pero se derivan de la intencionalidad de los agentes humanos.

He usado la forma terminológica «intrínseco» durante más de una década (Searle, 1980b), pero está sometida a ciertos malentendidos persistentes. En el habla ordinaria, «intrínseco» se opone a menudo a «relacional». De modo que la Luna tiene intrínsecamente masa, pero no es intrínsecamente un satélite. Sólo es un satélite en relación a la Tierra. En

este sentido de «intrínseco», quienes crean en estados intencionales con «contenido amplio», que es el contenido que involucra esencialmente relaciones con objetos exteriores a la mente, se verán obligados a negar que tales estados intencionales sean intrínsecos, porque son relacionales. No creo en la existencia del contenido amplio (véase Searle, 1983, cap. 7), de modo que el problema no se me plantea. Las distinciones que estoy haciendo en este momento son independientes del debate sobre el contenido amplio y estricto. De modo que sólo estoy estipulando que por «intencionalidad intrínseca» me refiero a la cosa real, en tanto que opuesta a las meras apariencias de la cosa (*como-si*), y en tanto que opuesta a las formas derivadas de intencionalidad como las oraciones, los retratos, etc. No es necesario aceptar mis objeciones al contenido amplio para aceptar las distinciones que estoy tratando de hacer.

Otro malentendido —que me resulta sorprendente— es suponer que al denominar «intrínsecas» a las instancias de la cosa real estoy implicando que son de algún modo misteriosas, inefables y más allá del alcance de la explicación filosófica o del estudio científico. Pero esto es absurdo. Tengo ahora muchos estados intencionales intrínsecos, por ejemplo, las ganas de ir al lavabo, el fuerte deseo de beber una cerveza fría, la experiencia visual de muchos botes en el lago. Todos estos son estados intencionales *intrínsecos*, en mi opinión, lo que sólo quiere decir que son la cosa real y no solamente algo más o menos parecido a la cosa real (*como-si*), o algo que es el resultado de los usos o actitudes de alguien más hacia la cosa (*derivados*).[3]

He observado muchos intentos de negar estas distinciones, pero es muy difícil tomárselos en serio. Quien piense que no hay diferencias de principio, podría considerar el siguiente fragmento de la revista *Pharmacology*.

> Una vez que el alimento ha atravesado el esfínter de la faringe, su movimiento es casi completamente involuntario, excepto por la expulsión final de las heces durante la defecación. *El tracto gastrointestinal es un órgano muy inteligente que percibe*, no sólo la presencia del alimento, sino también su composición química, su cantidad, su viscosidad y ajusta la frecuencia de mezcla y propulsión produciendo los patrones apropiados de contracciones. *Debido al enorme desarrollo de su capacidad de tomar decisiones*, la pared intestinal, constituida por finas capas de músculos, estructuras neuronales y las células paracrinas-endo-

3. Como ejemplo de este malentendido, véase P. M. y P. S. Churchland (1983).

> crinas, es *denominada a menudo el cerebro intestinal* (Sarna y Otterson, 1988, las cursivas son mías.)[4]

Se trata de un caso claro de intencionalidad *como-si* del «cerebro intestinal». ¿Cree alguien que no hay una diferencia de principio entre el cerebro intestinal y el cerebro cerebral? He oído decir que ambos tipos de caso son el mismo; de que todo es asunto de tomar una determinada «postura intencional» en relación a un sistema. Pero, simplemente, tratemos de suponer en la vida real que la «percepción» y la «toma de decisión» del cerebro intestinal no son diferentes de las del cerebro real.

El ejemplo anterior revela, entre otras cosas, que cualquier intento de negar la distinción entre intencionalidad intrínseca e intencionalidad *como-si* se enfrenta a una reducción al absurdo general. Si negamos la distinción, nos encontramos con que todo el universo tiene intencionalidad. Todo en el universo sigue las leyes de la naturaleza, y, por esa razón, todo se comporta con cierto grado de regularidad, y, por esa razón, todo se comporta *como si* estuviera siguiendo una regla, tratando de ejecutar un proyecto, actuando de acuerdo con ciertos deseos, etc. Supongamos por ejemplo, que dejo caer una piedra. La piedra *trata* de alcanzar el centro de la Tierra porque *desea* alcanzar el centro de la Tierra, y, al hacerlo así, *sigue la regla* $V = 1/2gt^2$. El precio de negar la distinción entre intencionalidad intrínseca y *como-si* es, en pocas palabras, el absurdo, porque hace que cualquier cosa del universo tenga mente.

Sin ninguna duda hay casos marginales. No estamos seguros de qué decir, por ejemplo, respecto a las moscas o los grillos. Y, sin ninguna duda, incluso en algunos casos humanos podríamos sentirnos perplejos respecto a si debemos tomar la adscripción de intencionalidad en un sentido literal o metafórico. Pero los casos marginales no alteran la distinción entre el tipo de hechos que corresponden a las adscripciones de intencionalidad intrínseca y aquellas correspondientes a las adscripciones metafóricas de intencionalidad *como-si*. No hay nada perverso, confuso o filosóficamente equivocado en las adscripciones metafóricas *como-si*. El único error es el de tomarlas literalmente.

Espero que las distinciones que he estado haciendo sean desesperadamente obvias. Sin embargo, he de informar, por así decir, desde el

4. Estoy en deuda con Dan Rudermann, por llamar mi atención sobre este artículo.

frente de batalla, que la negación de estas distinciones simples subyace en algunos de los mayores errores en la vida intelectual de nuestros días. Un patrón común de error es suponer que, porque podemos hacer adscripciones *como-si* de intencionalidad a sistemas que no tienen intencionalidad intrínseca, hemos descubierto de un modo u otro la naturaleza de la intencionalidad.[5]

5. Véase, por ejemplo, Dennett (1987).

4. LA CONCIENCIA Y SU LUGAR EN LA NATURALEZA

I. LA CONCIENCIA Y LA IMAGEN «CIENTÍFICA» DEL MUNDO

Como con la mayoría de las palabras, no es posible proporcionar una definición de «conciencia» en términos de condiciones necesarias y suficientes, ni es posible definirla al modo aristótelico de género y diferencia específica. Sin embargo, aunque no podamos proporcionar una definición verbal no circular, es esencial para mí que pueda decir lo que quiero decir con esa noción, porque es confundida a menudo con otras. Por ejemplo, por razones tanto etimológicas como de uso, «conciencia» es confundida con «autoconciencia» y «cognición».

Lo que quiero decir por «conciencia» puede ser ilustrado mejor por medio de ejemplos. Cuando me despierto, después de dormir sin haber soñado, paso a estar consciente, un estado que continúa tanto tiempo como estoy despierto. Cuando voy a dormir, o me ponen bajo una anestesia general, o muero, mis estados conscientes cesan. Si sueño mientras duermo, adquiero de nuevo conciencia, aunque las formas de conciencia en el sueño son, en general, de un nivel de intensidad mucho menor que la conciencia ordinaria mientras estamos despiertos. La conciencia puede variar en grado, incluso durante las horas de vigilia, como, por ejemplo, cuando pasamos de estar completamente despiertos y atentos a estar adormecidos y relajados, o, simplemente, cansados y poco atentos. Algunas personas introducen sustancias químicas en sus cerebros con el propósito de producir estados de conciencia alterados, pero, incluso sin ayuda química, es posible distinguir en la vida ordinaria entre diferentes grados y formas de conciencia. La conciencia es como un mecanismo de encendido y apagado: un sistema es consciente o no lo

es. Pero, una vez consciente, el sistema admite diferentes intensidades: hay grados diferentes de conciencia.

Un término casi sinónimo de «conciencia», en mi opinión, es el de «aprehensión» (*awareness*), pero no creo que sean exactamente equivalentes en significado porque «aprehensión» está más estrechamente vinculado a la cognición, al conocimiento, de lo que lo está la noción general de conciencia. Además, parece posible que podamos aceptarr que, en ciertos casos, alguien aprehenda o capte algo de un modo inconsciente (cf. Weiskrantz *et al.*, 1974). También vale la pena poner de manifiesto que no hay nada en mi análisis de la conciencia que implique la *auto*conciencia. Discutiré más adelante (en el capítulo 6) la conexión entre conciencia y autoconciencia.

Algunos filósofos (por ejemplo, Block en «Two Concepts of Consciousness») pretenden que hay un sentido de esta palabra que no implica sentir nada en absoluto, un sentido en el que un completo zombi podría ser «consciente». No conozco tal sentido, pero, en cualquier caso, no es ese el sentido en el que estoy utilizando la palabra.

Los estados conscientes siempre tienen contenido. No es posible ser sólo consciente, debe haber una respuesta a la cuestión «¿De qué eres consciente?». Pero el «de» de «consciente de» no siempre es el «de» de la intencionalidad. Si soy consciente de un golpe en la puerta, mi estado consciente es intencional, porque hace referencia a algo más allá de él mismo, el golpe en la puerta. Si soy consciente de un dolor, el dolor no es intencional, porque no representa nada más allá de sí mismo.[1]

El propósito principal de este capítulo es el de localizar la conciencia dentro de nuestra concepción «científica» global del mundo. La razón para subrayar la conciencia en un análisis de la mente es simplemente que se trata de la noción mental central. De un modo u otro, todas las demás nociones mentales —como intencionalidad, subjetividad, causalidad mental, inteligencia, etc.— sólo pueden ser entendidas completamente como *mentales* a través de sus relaciones con la conciencia (sobre esta cuestión, véase el capítulo 7). Dado que, en cual-

1. Hay que hacer una matización al respecto. El sentido de la localización corporal tiene intencionalidad, porque se refiere a una parte del cuerpo. Este aspecto de los dolores es intencional, porque tiene condiciones de satisfacción. En el caso de un miembro fantasma, por ejemplo, es posible estar equivocado y la posibilidad de error es, como mínimo, una buena pista de que el fenómeno es intencional.

quier momento de nuestras vidas, sólo una pequeña fracción de nuestros estados mentales es consciente, puede parecer paradójico pensar en la conciencia como la noción mental central, pero intento, en el curso de este libro, solucionar esta apariencia de paradoja. Una vez que hayamos localizado el lugar de la conciencia en nuestra visión general del mundo, podremos ver que las teorías materialistas de la mente que discutimos en el capítulo 2 son tan profundamente anticientíficas como el dualismo que pensaban que estaban atacando.

Descubriremos que, cuando tratemos de enunciar los hechos, la presión sobre las categorías y la terminología tradicionales se convierte en casi insoportable y comienzan a resquebrajarse por la tensión a que se ven sometidas. Lo que digo puede parecer casi autocontradictorio: por una parte, defenderé que la conciencia es sólo un rasgo biológico ordinario del mundo, pero también intentaré mostrar por qué encontramos casi inconcebible que deba ser de ese modo.

Nuestra concepción contemporánea del mundo comenzó a desarrollarse durante el siglo XVII, y su desarrollo continúa directamente a lo largo del siglo XX. Históricamente, una de las claves para este desarrollo fue la exclusión de la conciencia del tema objeto de la ciencia por parte de Descartes, Galileo y otros en el XVII. Según la concepción cartesiana, las ciencias naturales genuinas excluían la «mente», *res cogitans*, y se preocupaban sólo de la materia, *res extensa*. La separación de mente y materia fue una herramienta heurística útil en el siglo XVII, una herramienta que facilitó gran parte del progreso que tuvo lugar en las ciencias. Sin embargo, la separación es confusa filosóficamente, y, en el siglo XX, se había convertido en un obstáculo enorme para la comprensión científica del lugar de la mente en el mundo natural. Uno de los propósitos principales de este libro es el de tratar de apartar ese obstáculo, el de devolver la conciencia al tema que es el objeto de la ciencia en tanto que fenómeno biológico como cualquier otro. Para ello, necesitamos responder las objeciones dualistas de los cartesianos contemporáneos.

No hace falta decir que nuestra concepción «científica» del mundo es extremadamente compleja e incluye todas nuestras teorías generalmente aceptadas sobre qué tipo de lugar es el universo y cómo funciona. Es decir, incluye teorías que van de la mecánica cuántica y la teoría de la relatividad a la teoría geológica de la tectónica de placas y la teoría del ADN sobre la transmisión de la herencia. En nuestros días, por ejemplo, incluye la creencia en los agujeros negros, la teoría de las en-

fermedades causadas por gérmenes y la explicación heliocéntrica del sistema solar. Algunos aspectos de esta concepción del mundo son meras sugerencias, otros están bien establecidos. Al menos, dos aspectos son tan fundamentales y están tan bien establecidos que ya no son opciones abiertas para los ciudadanos razonablemente bien educados de esta época; de hecho, son en gran parte constitutivos de la moderna concepción del mundo. Se trata de la teoría atómica de la materia y la teoría evolucionista en biología. Por supuesto, como cualquier otra teoría, podrán ser refutadas por investigaciones posteriores, pero, por el momento, la evidencia a su favor es tan abrumadora que no parecen estar a la espera de que alguien las refute. Para situar la conciencia dentro de nuestra concepción del mundo, debemos situarla respecto a estas dos teorías.

De acuerdo con la teoría atómica de la materia, el universo consiste por completo en fenómenos físicos extremadamente pequeños a los que, por conveniencia, aunque no sea del todo apropiado, denominamos «partículas». Todas las cosas grandes y de tamaño medio que hay en el mundo, como los planetas, las galaxias, los coches y los abrigos, están hechas de pequeñas partículas que, a su vez, están compuestas de partículas más pequeñas todavía, hasta que finalmente alcanzamos el nivel de las moléculas, compuestas de átomos, que, a su vez, se componen de partículas subatómicas. Son ejemplos de partículas, los electrones, los átomos de hidrógeno y las moléculas de agua. Como ilustran estos ejemplos, las partículas mayores están compuestas de partículas más pequeñas; y todavía hay una enorme incertidumbre y mucha discusión sobre la identificación de las partículas más pequeñas de todas. No estamos del todo satisfechos con el uso de la palabra «partícula» por dos razones como mínimo. En primer lugar, parece más exacto describir las más básicas de esas entidades como puntos de masa/energía más bien que como entidades espaciales extendidas. Y, en segundo lugar, más radicalmente, la mecánica cuántica nos dice que, en cuanto no son medidas o interferidas de algún modo, las «partículas» como los electrones se comportan más como ondas que como partículas. Sin embargo, por motivos de conveniencia, seguiré utilizando la palabra «partícula».

Las partículas, como ilustraban los ejemplos anteriores, se organizan en *sistemas* mayores. Sería un tanto complicado tratar de definir la noción de sistema, pero la idea intuitiva más simple es la de que los sistemas son conjuntos de partículas en los que los límites espacio-tem-

porales del sistema están fijados por las relaciones causales. Así, una gota de lluvia es un sistema, pero también lo es un glaciar. Los bebés, los elefantes y las cadenas de montañas son también ejemplos de sistemas. A partir de esos ejemplos, resultará obvio que los sistemas pueden contener subsistemas.

Es esencial al aparato explicativo de la teoría atómica no sólo la idea de que los sistemas grandes están hechos de sistemas pequeños, sino también que muchos rasgos de los grandes pueden ser *causalmente explicados* por la conducta de los pequeños. Esta concepción de la explicación nos da la posibilidad, en realidad nos impone el requisito, de que muchos tipos de macrofenómenos sean explicables en términos de microfenómenos. Y esto, a su vez, tiene la consecuencia de que habrá niveles diferentes de explicación del mismo fenómeno, dependiendo de si vamos de derecha a izquierda —de lo macro a lo macro o de lo micro a lo micro— o de abajo arriba —de lo micro a lo macro. Podemos ilustrar estos niveles con un simple ejemplo. Supongamos que deseo explicar por qué hierve este cazo de agua. Una explicación de izquierda a derecha, de lo macro a lo macro, sería la de que puse el cazo en la cocina y encendí el fuego bajo él. Denomino esta explicación «izquierda-derecha» porque menciona un suceso anterior para explicar uno posterior,[2] y la denomino «macro-macro» porque tanto el *explanans* como el *explanandum* están al macronivel. Otra explicación, de abajo arriba, micro-macro, sería la de que el agua está hirviendo porque la energía cinética transmitida por la oxidación de los hidrocarbonos a las moléculas de H_2O ha causado que se muevan tan rápidamente que la presión interna de los movimientos moleculares iguala la presión del aire exterior, la cual, a su vez, se explica por el movimiento de las moléculas de las que está compuesto el aire exterior. Denomino esta explicación «de abajo arriba, micro-macro» porque explica los rasgos y la conducta superficial de los macrofenómenos en términos de microfenómenos de nivel inferior. No quiero implicar que estos sean los únicos niveles de explicación posibles. También hay explicaciones de izquierda a derecha, micro-micro, y, en el seno de cada nivel micro o macro, pueden realizarse subdivisiones adicionales.

Esta es, pues, una de las lecciones más importantes de la teoría atómica: muchos rasgos de las cosas grandes son explicados por la conducta

2. La metáfora «izquierda-derecha» deriva, por supuesto, de la convención arbitraria de las lenguas europeas de escribir de izquierda a derecha.

de las cosas pequeñas. Consideramos la teoría de la enfermedad causada por gérmenes o la teoría de la transmisión genética por el ADN como hitos tan fundamentales precisamente porque se ajustan a este modelo. Si alguien tuviera una explicación de las enfermedades en términos de los movimientos planetarios, nunca la aceptaríamos como una explicación completa, incluso si funcionara para curas y diagnósticos, hasta que entendiéramos cómo las macrocausas y macroefectos a nivel de los planetas y los síntomas están fundamentados en estructuras causales de abajo arriba, micro-macro.

Añadamos los principios de la biología evolucionista a estas nociones elementales de la teoría atómica. Durante largos períodos de tiempo, ciertos *tipos* de sistemas vivos evolucionan de ciertas maneras muy especiales. En nuestra pequeña Tierra, los tipos de sistemas en cuestión contienen invariablemente moléculas basadas en el carbono, y en ellas abundan el oxígeno, el hidrógeno y el nitrógeno. Las maneras en las que evolucionan son complicadas, pero el proceso básico es que las instancias particulares de los tipos llevan a la existencia a instancias similares. Así, cuando las instancias originales han sido destruidas, el tipo de patrón por ellas ejemplificado continúa en otras instancias y continúa siendo replicado a medida que generaciones subsiguientes de instancias producen otras instancias. Variaciones en los rasgos superficiales, o fenotipos, de las instancias les proporcionan mayores o menores posibilidades de supervivencia, en relación a los ambientes específicos en que se encuentran. Las instancias que tienen mayores probabilidades de supervivencia en relación a su ambiente tendrán, por tanto, mayor probabilidad de producir otras instancias semejantes a ellas, instancias con el mismo genotipo. Y así es como el tipo evoluciona.

Parte del atractivo intelectual de la teoría de la evolución, complementada por la genética mendeliana y la del ADN, es que se ajusta al modelo explicativo que hemos derivado de la teoría atómica. Específicamente, la fundamentación de los mecanismos genéticos en la biología molecular permite diferentes niveles de explicación de los fenómenos biológicos correspondientes a los diferentes niveles de explicación que tenemos para los fenómenos físicos. En la biología evolucionista, hay típicamente dos niveles de explicación, un nivel «funcional» en el que explicamos la supervivencia de la especie en términos de «adecuación inclusiva», que depende de los rasgos fenotípicos poseídos por los miembros de la especie, y un nivel «causal» en el que explicamos los me-

canismos causales por medio de los cuales los rasgos en cuestión relacionan efectivamente los organismos con su medio ambiente. Podemos ilustrarlo con un ejemplo simple. ¿Por qué las plantas verdes giran sus hojas hacia el Sol? La explicación funcional dice:[3] este rasgo tiene valor de supervivencia. Al incrementar la capacidad de la planta de realizar la fotosíntesis, incrementa la capacidad de la planta de sobrevivir y reproducirse. La planta no gira hacia el Sol para sobrevivir; más bien, la planta tiende a sobrevivir porque está predispuesta a girar hacia el Sol en cualquier caso. La explicación causal afirma: la estructura bioquímica de la planta en tanto que determinada por su equipamiento genético causa que secrete la hormona del crecimiento, la auxina, y las diferentes concentraciones de auxina causan a su vez que las hojas giren en dirección a la fuente de la luz.

Si ponemos juntos estos dos niveles de explicación, tenemos el resultado siguiente: el genotipo sobrevive y se reproduce porque el fenotipo, en tanto que producido por la interacción del genotipo con el medio ambiente, tiene un valor de supervivencia en relación al medio ambiente. En forma muy breve, estos son los mecanismos de la selección natural.

Los productos del proceso de la evolución, los organismos, están hechos de subsistemas denominados «células», y algunos de estos organismos desarrollan subsistemas de células nerviosas, que concebimos como «sistemas nerviosos». Además, y este es el punto crucial, algunos sistemas nerviosos extremadamente complejos son capaces de causar y mantener procesos y estados conscientes. No conocemos los detalles de cómo los cerebros causan la conciencia, pero sabemos que es un hecho que tal cosa sucede en los cerebros humanos, y tenemos evidencia abrumadora de que también ocurre en el cerebro de muchas especies de animales (Griffin, 1981). No sabemos por el momento cuál es el punto más bajo en la escala de la evolución hasta el que se extiende la conciencia.

Una idea básica en nuestra concepción del mundo es que los seres humanos y otros animales superiores son parte del orden biológico como cualquier otro organismo. Los seres humanos son una continuación del resto de la naturaleza. Pero, si es así, las características biológicas

3. El término «funcional», de algún modo, lleva a confusión porque el nivel funcional también es causal, pero es frecuente en biología hablar de los dos tipos de explicación causal como «funcional» y «causal». Independientemente de cómo la describamos, la distinción es importante y la utilizaré más adelante, en el capítulo 10.

gicamente específicas de estos animales —como el hecho de que posean un sistema rico de conciencia y mayor inteligencia, su capacidad para el lenguaje, su capacidad para discriminaciones perceptivas extremadamente refinadas, su capacidad para el pensamiento racional, etc.— son fenómenos biológicos como cualquier otro fenómeno biológico. Además, todo esto son rasgos del fenotipo. Son tan resultado de la evolución biológica como cualquier otro fenotipo. *En pocas palabras, la conciencia es un rasgo biológico de los cerebros humanos y de ciertos animales. Está causada por procesos neurobiológicos y es una parte del orden biológico natural como cualquier otro rasgo biológico, como lo son la fotosíntesis, la digestión o la mitosis.* Este principio es el primer estadio para la comprensión del lugar de la conciencia en el seno de nuestra concepción del mundo.[4] La tesis de este capítulo hasta el momento ha sido la de que, una vez que vemos que las teorías atómicas y evolucionistas son centrales para la concepción científica del mundo, la conciencia encuentra su lugar naturalmente como un rasgo fenotípico de ciertos tipos de organismos con sistemas nerviosos altamente desarrollados. En este capítulo, no trato de defender esta concepción del mundo. De hecho, muchos pensadores cuyas opiniones respeto, el caso más notable es el de Wittgenstein, la consideran de algún modo repudiable y degradante. Les parece que no hay lugar en ella —al menos un lugar que no sea subsidiario— para la religión, el arte, el misticismo y los valores «espirituales» en general. Pero, guste o no guste, es la concepción del mundo que tenemos. Dado lo que sabemos sobre los detalles del mundo —sobre cosas tales como la posición de los elementos en la tabla periódica, el número de cromosomas en las células

4. A veces, mis puntos de vista encuentran resistencia a causa de una concepción equivocada de las relaciones entre causalidad e identidad. U. T. Place (1988), por ejemplo, escribe: «De acuerdo con Searle, los estados mentales son, a la vez, idénticos a, y causalmente dependientes de, los correspondientes estados cerebrales. Mi posición es que no es posible nadar y guardar la ropa. O bien los estados mentales son idénticos a los estados cerebrales o unos dependen causalmente de los otros. No pueden ser ambas cosas» (p. 209).

Place piensa en casos como «Estas huellas pueden depender causalmente de los zapatos del ladrón, pero no pueden ser, a la vez, idénticas a esos zapatos». Pero ¿qué pasa con «El estado líquido de este agua puede ser causalmente dependiente de la conducta de las moléculas, y también puede ser un rasgo del sistema que está compuesto por las moléculas»? Me parece igualmente obvio que mi presente estado de conciencia está causado por la conducta neuronal de mi cerebro y que ese mismo estado es sólo un rasgo de nivel superior del cerebro. Si esto quiere decir nadar y guardar la ropa, nademos.

de las diferentes especies y la naturaleza de los vínculos químicos—, esta concepción del mundo no es una opción abierta al debate. No compite con otras muchas concepciones alternativas del mundo. Nuestro problema no es que hemos fracasado de algún modo a la hora de encontrar una prueba convincente de la existencia de Dios, o que la hipótesis de la vida después de la muerte es algo seriamente dudoso, es más bien que en nuestras más profundas reflexiones no podemos tomar en serio tales opiniones. Cuando nos encontramos con alguien que cree tales cosas, podemos envidiar la tranquilidad y la seguridad que pretende obtener de esas creencias, pero en el fondo continuamos convencidos de que o bien no se ha enterado de lo que pasa, o bien es prisionero de la fe. Estamos convencidos de que, de algún modo, debe dividir su mente en compartimientos estancos para creer cosas como esas. Cuando di conferencias sobre el problema mente-cuerpo en la India y algunos miembros de mi auditorio me aseguraron que mis puntos de vista deberían estar equivocados porque ellos en persona habían existido en vidas anteriores como ranas o elefantes, etc., no pensé «he aquí evidencia para una concepción alternativa del mundo», ni siquiera «quién sabe, a lo mejor están en lo cierto». Mi falta de sensibilidad era mucho más que provincianismo cultural: dado lo que sé sobre cómo funciona el mundo, no podía considerar sus puntos de vista como candidatos serios a la verdad.

Una vez que alguien acepta nuestra cosmovisión, el único obstáculo para garantizar a la conciencia su estatus como un rasgo biológico de los organismos es el desprestigiado supuesto materialista-dualista de que el carácter «mental» de la conciencia convierte en imposible el que sea una propiedad «física».

Sólo he discutido la relación de la conciencia con los sistemas vivos basados en el carbono del tipo de los que tenemos en nuestra Tierra, pero, por supuesto, no podemos excluir la posibilidad de que la conciencia pueda haber evolucionado en otros planetas de otros sistemas solares en otras partes del universo. Dado el tamaño enorme del universo, sería sorprendente en términos estadísticos que fuéramos nosotros los únicos portadores de la conciencia. Además, no deseamos excluir la posibilidad de que la conciencia pudiera haber evolucionado en sistemas que, en vez de estar basados en el carbono, usaran algún tipo de química completamente diferente. Por todo lo que sabemos hoy en día, podría no existir ningún obstáculo teórico al desarrollo de la conciencia en sistemas hechos de otros elementos. Por el momento, es-

tamos muy lejos de tener una teoría adecuada de la neurofisiología de la conciencia, pero, hasta que la tengamos, debemos mantener una actitud abierta respecto a sus bases químicas posibles. Mi propia sospecha es que es probable que la neurobiología de la conciencia se pruebe, al menos, tan restrictiva como la bioquímica de la digestión. Hay variedades diversas de digestión, pero no cualquier cosa puede ser digerida por cualquier cosa. De un modo semejante, me parece que es probable que descubramos que, aunque puedan existir variedades de conciencia bioquímicamente diferentes, no todo vale.

Además, dado que la conciencia está causada por completo por la conducta de fenómenos biológicos de nivel inferior, sería posible en principio producirla artificialmente duplicando los poderes causales del cerebro en una situación de laboratorio. Sabemos que muchos fenómenos biológicos se han creado de un modo artificial. Podemos sintetizar ciertos compuestos orgánicos, e incluso crear artificialmente ciertos procesos como la fotosíntesis. Si podemos crear la fotosíntesis de un modo artificial, ¿por qué no la conciencia también? En el caso de la fotosíntesis, la forma artificial del fenómeno se creó duplicando de hecho en el laboratorio los procesos químicos. De un modo semejante, si alguien tratara de crear la conciencia de un modo artificial, la manera natural de hacerlo sería intentar duplicar la base neurofisiológica que la conciencia tiene de hecho en organismos como nosotros. Dado que, en el momento presente, no sabemos exactamente cuál es la base neurobiológica, las perspectivas para tal «inteligencia artificial» son muy remotas. Además, como sugería anteriormente, podría ser posible producir conciencia usando un tipo de química totalmente diferente al que nuestro cerebro usa de hecho. Sin embargo, una cosa que sabemos incluso antes de comenzar la investigación es que *cualquier sistema capaz de causar conciencia debe ser un sistema capaz de duplicar los poderes causales del cerebro.* Si, por ejemplo, se hace con piezas de silicio en vez de neuronas, debe ser porque la química de las piezas de silicio es capaz de duplicar los específicos poderes causales de las neuronas para causar la conciencia. Es una consecuencia lógica trivial del hecho de que los cerebros causen conciencia el que cualquier otro sistema capaz de causar conciencia, usando mecanismos completamente diferentes, debería tener al menos el poder equivalente al que tienen los cerebros para hacerlo. (Compárese: los aviones no necesitan plumas para volar, pero tienen que compartir con los pájaros la capacidad causal de vencer la fuerza de la gravedad en la atmósfera terrestre).

Para resumir: nuestra imagen del mundo, aunque extremadamente complicada en detalle, proporciona una explicación más bien simple del modo de existencia de la conciencia. De acuerdo con la teoría atómica, el mundo está hecho de partículas. Esas partículas están organizadas en sistemas. Algunos de esos sistemas tienen vida, y esos tipos de sistemas vivos han evolucionado a lo largo de enormes períodos de tiempo. Entre ellos, algunos han desarrollado cerebros que son capaces de causar y mantener conciencia. La conciencia es, así, un rasgo biológico de ciertos organismos en exactamente el mismo sentido de «biológico» en el que la fotosíntesis, la mitosis, la digestión y la reproducción son rasgos biológicos de los organismos.

He intentado describir la posición de la conciencia en nuestra cosmovisión general de un modo muy simple, porque quiero que parezca absolutamente obvia. Cualquiera que haya tenido una educación «científica», aunque sea mínima, después de 1920 no puede encontrar nada controvertible en lo que acabo de decir. También vale la pena subrayar que todo esto se ha dicho sin la ayuda de las categorías cartesianas tradicionales. No se han discutido el dualismo, el monismo, el materialismo ni nada de ese tipo. Además, no ha habido ninguna mención a la «naturalización del contenido»; ya es algo completamente natural. Para repetirlo una vez más: la conciencia es un fenómeno biológico natural. La exclusión de la conciencia del mundo natural fue un recurso heurístico útil en el siglo XVII, porque permitió a los filósofos concentrarse en fenómenos que fueran medibles, objetivos y desprovistos de significado, es decir, libres de intencionalidad. Pero la exclusión estaba basada en una falsedad. Se basaba en la creencia falsa de que la conciencia no es parte del mundo natural. Esa única falsedad, más que cualquier otra, más incluso que la enorme dificultad de estudiar la conciencia con las herramientas científicas a nuestro alcance, nos ha impedido llegar a una comprensión de la conciencia.

II. SUBJETIVIDAD

Los estados y procesos mentales conscientes tienen un rasgo especial no poseído por otros fenómenos naturales, o sea, la subjetividad. Es ese rasgo de la conciencia el que hace su estudio tan recalcitrante a los métodos convencionales de la investigación biológica y psicológica, y una fuente de extrema perplejidad para el análisis filosófico. Hay

diferentes sentidos de «subjetividad», ninguno de ellos completamente claro, y necesito decir al menos un poco más para clarificar el sentido en el que afirmo que la conciencia es subjetiva.

Decimos a veces que los juicios son «subjetivos» cuando queremos decir que su verdad o falsedad no puede ser establecida «objetivamente», porque su verdad o falsedad no es simplemente un asunto de hecho, sino que depende de ciertas actitudes, sentimientos y puntos de vista de quienes realizan y escuchan el juicio. Un ejemplo de tal tipo de juicio podría ser «Van Gogh es un artista mejor que Matisse». En este sentido de «subjetividad», oponemos estos juicios subjetivos a juicios completamente objetivos, como el juicio «Matisse vivió en Niza durante el año 1917». Para tales juicios objetivos, podemos discernir qué tipos de hechos en el mundo los convierten en verdaderos o falsos, independientemente de las actitudes o sentimientos de alguien sobre ellos.

Ahora bien, este sentido en el que hablamos de juicios «subjetivos» y «objetivos» no es el sentido de «subjetivo» en el que hablo de la conciencia como algo subjetivo. En el sentido en el que uso el término, «subjetivo» se refiere a una categoría ontológica, no a un modo epistemológico. Consideremos, por ejemplo, el enunciado «Tengo dolor de espalda». Tal enunciado es completamente objetivo en el sentido de que lo convierte en verdadero la existencia de un hecho real y no depende de las actitudes u opiniones de los observadores. Sin embargo, el fenómeno mismo, el dolor real mismo, tiene un modo subjetivo de existencia, y es en ese sentido en el que digo que la conciencia es subjetiva.

¿Qué más puede decirse sobre este modo subjetivo de existencia? En primer lugar, es esencial ver que, como consecuencia de su subjetividad, el dolor no es igualmente accesible a todo observador. Podríamos decir que su existencia es una existencia para-la-primera-persona. Para que algo sea un dolor, debe ser un dolor de *alguien*, y serlo en un sentido más fuerte que el sentido en el que, por ejemplo, decimos que una pierna debe ser una pierna de alguien. Los trasplantes de pierna son posibles; en ese sentido, los trasplantes de dolor no lo son. Y lo que es verdad de los dolores es verdad de los estados conscientes en general. Todo estado consciente es siempre el estado consciente de *alguien*. Y, del mismo modo en que tengo una relación especial con mis estados conscientes, que no es como mi relación con los estados conscientes de los demás, ellos tienen una relación con sus estados conscientes que no

es igual a la relación que yo tengo con sus estados conscientes.[5] La subjetividad tiene la consecuencia adicional de que todas mis formas conscientes de intencionalidad que me dan información sobre el mundo independiente de mí siempre son desde un punto de vista especial. El mundo mismo no tiene ningún punto de vista, pero mi acceso al mundo a través de mis estados mentales conscientes siempre se da desde una perspectiva, desde mi punto de vista.

Sería difícil exagerar los efectos desastrosos que el no tratar adecuadamente la subjetividad de la conciencia ha tenido sobre la producción filosófica y psicológica del último medio siglo. De modos que no siempre son obvios a primera vista, gran parte de la bancarrota de la mayoría de los trabajos en filosofía de la mente y gran parte de la esterilidad de la psicología académica durante los últimos cincuenta años, durante todo el tiempo de mi vida intelectual, tiene su origen en el continuo fracaso a la hora de reconocer y tratar adecuadamente el hecho de que la ontología de lo mental es, irreductiblemente, una ontología de la primera persona. Hay muchas razones muy profundas, muchas de ellas localizadas en nuestra historia inconsciente, por las que encontramos difícil, si no imposible, aceptar la idea de que el mundo real, el mundo descrito por la física, la química y la biología, contenga un elemento ineliminablemente subjetivo. ¿Cómo podría ser tal cosa? ¿Cómo es posible que podamos formar una imagen coherente del mundo si el mundo contiene esas misteriosas entidades conscientes? Y, sin embargo, todos nosotros sabemos que estamos conscientes la mayor parte de nuestras vidas, y que las personas a nuestro alrededor también son conscientes. Y, a menos que estemos cegados por la mala filosofía o por algunas formas de psicología académica, no tenemos duda alguna de que los perros, los gatos, los monos y los niños pequeños son conscientes, y que su conciencia es algo tan subjetivo como la nuestra.

Describamos, pues, con un poco más de detalle la imagen del mundo que contiene la subjetividad como un elemento fundamental, e intentemos describir luego algunas de las dificultades que tenemos para aceptar esa imagen del mundo. Si concebimos el mundo como algo compuesto de partículas y esas partículas como algo organizado en sistemas, y algunos de esos sistemas como sistemas biológicos, y algunos

5. Este no es un argumento a favor del acceso privilegiado porque no hay ningún privilegio, ni ningún acceso. Diré algo más sobre este tema más adelante, en el presente capítulo.

de esos sistemas biológicos son conscientes, y la conciencia es esencialmente subjetiva, ¿qué es lo que se nos pide que imaginemos cuando imaginamos la subjetividad de la conciencia? Después de todo, todas esas otras cosas que imaginábamos —partículas, sistemas, organismos, etc.— eran completamente objetivas. En consecuencia, son accesibles de la misma manera a todos los observadores competentes. De modo que, ¿qué es lo que se nos pide que imaginemos si añadimos a este caldero metafísico algo que es irreductiblemente subjetivo?

De hecho, lo que se nos pide que «imaginemos» es simplemente el mundo que sabemos que existe. Sé, por ejemplo, que ahora estoy consciente y que ese estado consciente en el que estoy tiene la subjetividad a la que me he referido, y sé que un enorme número de otros organismos como yo mismo son asimismo conscientes y tienen estados subjetivos similares. En ese caso, ¿por qué parece que pido que imaginemos algo difícil o, en cierto sentido, contraintuitivo, cuando lo único que hago es recordar hechos que están ahí, frente a nosotros, desde hace mucho tiempo? Parte —pero sólo parte— de la respuesta tiene que ver con el hecho de que, con toda ingenuidad, en el parágrafo previo he invocado la palabra «observador». Cuandos se nos pide que nos formemos una cosmo*visión* o *imagen* del mundo, nos la formamos acudiendo a modelos visuales. Literalmente, tendemos a formar una imagen de la realidad como algo consistente en pequeños trocitos de materia, las «partículas», y luego las imaginamos organizadas en sistemas, de nuevo con rasgos visibles a simple vista. Pero, cuando visualizamos el mundo con este ojo interior, no podemos ver la conciencia. De hecho, es la misma subjetividad de la conciencia la que la hace invisible de modo relevante. *Si tratamos de dibujar la imagen de la conciencia de otro, acabamos dibujando a la otra persona* (quizás con una especie de globo saliendo de su cabeza). *Si tratamos de dibujar nuestra propia conciencia, acabamos dibujando aquello de lo que somos conscientes.* Si la conciencia es la base epistémica más fundamental para capturar la realidad, no podemos alcanzar de ese modo la realidad de la conciencia. (Formulación alternativa: no podemos capturar la realidad de la conciencia del modo en que, usando la conciencia, podemos capturar la realidad de otros fenómenos.)

Es importante examinar despacio estos problemas y no pasar por ellos apresuradamente, como se hace habitualmente. De modo que pondré una marcha lenta y me desplazaré paso a paso. Si intento observar la conciencia de otro, lo que observo no es su subjetividad, sino, simple-

mente, su conducta consciente, su estructura y las relaciones causales entre estructura y conducta. Además, observo las relaciones causales entre estructura y conducta, por una parte, y el contexto que le afecta (y que también es afectado por él), por otra. De manera que no hay ningún modo en que pueda observar la conciencia de otro como tal conciencia; más bien lo que observo es él, su conducta y las relaciones entre él, la conducta, la estructura y el contexto. ¿Qué pasa entonces con mis propios sucesos interiores? ¿No los puedo observar? El mismo hecho de la subjetividad, que tratábamos de observar, convierte tal observación en algo imposible. ¿Por qué? Porque donde está implicada la subjetividad, no hay distinción entre la observación y la cosa observada, entre la percepción y la cosa percibida. El modelo de la visión funciona bajo el presupuesto de que hay una distinción entre la cosa vista y la visión de ella. Pero no hay modo alguno de establecer esa diferencia para la «introspección». Cualquier introspección que tengo de mi estado mental consciente es ella misma ese estado consciente. Lo que no es decir que mis fenómenos mentales conscientes no se presenten en muchos niveles y variedades diferentes —más adelante, tendremos ocasión de examinar algunos de ellos en detalle. Es simplemente decir que el modo habitual de observación no funciona para la subjetividad consciente. No funciona para la conciencia de otras personas, y no funciona para la conciencia propia. Por ese motivo, la idea de que pudiera existir un método especial para investigar la conciencia, o sea la introspección, que se supone que es una suerte de observación interior, estaba condenada al fracaso desde el principio, y no es sorprendente que la psicología introspectiva acabara por fracasar.

Encontramos difícil explicar satisfactoriamente la subjetividad, no sólo por haber sido educados en una ideología que dice que, en último término, la realidad ha de ser completamente objetiva, sino porque nuestra idea de una realidad objetivamente observable presupone la noción de observación que es, en sí misma, ineliminablemente subjetiva, y que no puede convertirse en el objeto de la observación, como sí pueden serlo los objetos y estados de cosas objetivamente existentes. En pocas palabras, no hay manera en que podamos figurar pictóricamente la subjetividad como parte de nuestra cosmovisión porque, por así decirlo, la subjetividad en cuestión es la actividad de figuración pictórica. La solución no es la de intentar desarrollar un modo especial de representación pictórica, una suerte de superintrospección, si no más bien la de abandonar en este punto las figuraciones pictóricas y limitarnos a reco-

nocer los hechos. Los hechos son que los procesos biológicos producen fenómenos mentales conscientes, y que estos son irreductiblemente subjetivos.

Los filósofos han inventado otra metáfora para describir ciertos rasgos de la subjetividad que parece incluso más confusa que la metáfora de sentido común de la introspección: el «acceso privilegiado». Sentimos la tentación de sustituir la metáfora *visual* de la introspección por la metáfora *espacial* del acceso privilegiado, un modelo que sugiere que la conciencia es como una habitación privada en la que sólo nosotros estamos autorizados a entrar. Sólo yo puedo penetrar en el interior del espacio de mi propia conciencia. Pero esta metáfora tampoco funciona, porque, para que haya algo a lo que tengo acceso privilegiado, yo debería ser distinto al espacio en el que entro. Pero, exactamente del mismo modo en que la metáfora de la introspección se venía abajo cuando la única cosa a observar era la observación misma, la metáfora del espacio privado interior se rompe cuando comprendemos que no hay nada como un espacio en el que yo pueda entrar, porque no puedo establecer las distinciones necesarias entre los tres elementos: yo mismo, el acto de entrar y el espacio en el que se supone que entro.

Podríamos resumir estas consideraciones diciendo que nuestro modelo moderno de la realidad y de la relación entre realidad y observación no puede dar cuenta del fenómeno de la subjetividad. El modelo es el de observadores objetivos (en sentido epistémico) observando una realidad existente objetivamente (en sentido ontológico). Pero, en este modelo, no hay manera alguna de observar el mismo acto de observación. Dado que el acto de observación es el acceso subjetivo (sentido ontológico) a la realidad objetiva. Aunque puedo observar fácilmente a otra persona, no puedo observar su *subjetividad*. Y, mucho peor, no puedo *observar* mi propia subjetividad, dado que cualquier observación que pudiera procurarme sería, ella misma, aquello que se suponía que era observado. La idea de que hay una observación de la realidad es precisamente la idea de representaciones (ontológicamente) subjetivas de la realidad. La ontología de la observación —como opuesta a su epistemología— es precisamente la ontología de la subjetividad. La observación es siempre la observación de alguien; es, en general, la conciencia; siempre se da desde un punto de vista; tiene un sentimiento subjetivo asociado; etc.

Deseo dejar claro lo que estoy diciendo y lo que no estoy diciendo. No estoy realizando la vieja y confusa observación de qué hay una pa-

radoja de autorreferencia en el estudio de la subjetividad. Tales paradojas no me preocupan en absoluto. Podemos utilizar el ojo para estudiar el ojo, el cerebro para estudiar el cerebro, la conciencia para estudiar la conciencia, el lenguaje para estudiar el lenguaje, la observación para estudiar la observación y la subjetividad para estudiar la subjetividad. No hay ningún problema en ello. Lo importante es más bien que, a causa de la ontología de la subjetividad, nuestros modelos de «estudio», modelos que descansán en la distinción entre la observación y la cosa observada, no funcionan para la subjetividad misma.

Hay un sentido, pues, en el que nos es difícil concebir la subjetividad. Dado nuestro concepto de cómo debería ser esta realidad y cómo sería saber cómo es realidad, parece inconcebible que deba existir algo irreductiblemente subjetivo en el universo. Y, sin embargo, todos sabemos que la subjetividad existe.

Espero que podamos ver un poco más claramente lo que sucede si intentamos describir el universo dejando a un lado la subjetividad. Supongamos que insistimos en dar una explicación del mundo que sea completamente objetiva, no sólo en el sentido epistémico de que sus pretensiones sean comprobables independientemente, sino en el sentido ontológico de que los fenómenos que describa tengan una existencia independiente de cualquier forma de subjetividad. Una vez se adopta esta estrategia (la estrategia principal de la filosofía de la mente durante los últimos cincuenta años), es imposible describir la conciencia, porque es literalmente imposible el reconocimiento de la subjetividad de la conciencia. Los ejemplos son tan numerosos que no pueden mencionarse, pero citaré dos autores que se enfrentan explícitamente al problema de la conciencia. Armstrong (1980) elimina tácitamente la subjetividad al tratar la conciencia como una mera capacidad para realizar discriminaciones sobre los propios estados interiores y Changeux, el neurobiólogo francés, define la conciencia simplemente como «un sistema global regulatorio que trata de los objetos mentales y de las computaciones que usan esos objetos» (1985, p. 145). Ambos análisis presuponen una concepción de tercera persona de la realidad, una concepción de la realidad que no es sólo epistemológicamente objetiva, sino ontológicamente objetiva también; y tal realidad no tiene lugar alguno para la conciencia, porque no tiene lugar para la subjetividad ontológica.

III. La conciencia y el problema mente-cuerpo

He dicho repetidamente que creo que el problema mente-cuerpo tiene una solución más bien simple, al menos en términos generales, y que los únicos obstáculos para tener una comprensión completa de las relaciones mente-cuerpo son nuestro prejuicio filosófico, al suponer que lo mental y lo físico son dos ámbitos distintos, y nuestra ignorancia sobre el funcionamiento del cerebro. Si tuviéramos una ciencia adecuada del cerebro, una explicación del cerebro que proporcionara explicaciones causales de la conciencia en todas sus formas y variedades, y si superáramos nuestros errores conceptuales, no quedaría nada del problema mente-cuerpo. Sin embargo, la posibilidad de una solución al problema mente-cuerpo ha sido poderosamente cuestionada a lo largo de los años por los escritos de Thomas Nagel (1974, 1986). Argumenta del modo siguiente: por el momento, simplemente carecemos del aparato conceptual para siquiera concebir una solución al problema mente-cuerpo. La razón es la siguiente: las explicaciones en las ciencias naturales tienen un tipo de necesidad causal. Entendemos, por ejemplo, cómo la conducta de las moléculas de H_2O causa que el agua adopte la forma líquida porque vemos que el estado líquido es una consecuencia necesaria de la conducta de las moléculas. La teoría molecular hace algo más que mostrar que los sistemas de moléculas de H_2O serán líquidos bajo ciertas circunstancias; muestra más bien que el sistema *ha de adoptar* la forma líquida. Dado que comprendemos la física en cuestión, es inconcebible que las moléculas se comporten del modo relevante sin que el agua adopte la forma líquida. En pocas palabras, Nagel argumenta que las explicaciones en ciencia implican necesidad, y que la necesidad implica la inconcebibilidad de lo opuesto.

Ahora bien, según Nagel, no podemos alcanzar ese tipo de necesidad para la relación entre la materia y la conciencia. Ninguna posible descripción de la conducta neuronal explicaría por qué, dada esa conducta, *hemos de tener dolor*, por ejemplo. Ninguna explicación puede dar cuenta, por ejemplo, de por qué el dolor era una consecuencia necesaria de ciertos tipos de actividad neuronal. La prueba de que la explicación no nos proporciona necesidad causal es que siempre podemos concebir lo opuesto. Siempre podemos concebir un estado de cosas en el que lo neurofisiológico se comporta de la manera que queramos sin que, a pesar de ello, el sistema tenga dolor. Si la explicación científica adecuada implica necesidad y la necesidad implica la inconcebibilidad de lo

opuesto, entonces, por contraposición, la concebibilidad de lo opuesto implica que no tenemos necesidad, lo que, a su vez, implica que no tenemos una explicación. La desesperante conclusión de Nagel es la de que necesitaríamos una reestructuración enorme de nuestro sistema conceptual para poder resolver alguna vez el problema mente-cuerpo.

Este argumento no me convence. En primer lugar, debemos advertir que no todas las explicaciones en ciencia tienen el tipo de necesidad que encontramos en la relación entre el movimiento molecular y la liquidez. Por ejemplo, la ley del inverso del cuadrado de la distancia es una explicación de la gravedad, pero no muestra que los cuerpos *hayan de tener* atracción gravitatoria. En segundo lugar, la aparente «necesidad» de cualquier explicación científica puede ser sólo una función del hecho de que encontramos la explicación tan convincente que no podemos, por ejemplo, concebir que las moléculas se mueven de un modo particular y que el H_2O no es líquido. Una persona en los remotos tiempos de la Edad Media podría no haber encontrado la explicación un asunto de «necesidad». El «misterio» de la conciencia hoy en día está en, aproximadamente, la misma situación en que estaba el misterio de la vida antes del desarrollo de la biología molecular o el misterio del electromagnetismo antes de las ecuaciones Clerk-Maxwell. Parece misterioso porque no sabemos cómo funciona el sistema de la neurofisiología/conciencia, y un conocimiento adecuado de cómo lo hace eliminaría el misterio. Además, la pretensión de que siempre podemos concebir la posibilidad de que ciertos estados cerebrales *pudieran no* causar los estados conscientes apropiados podría depender sólo de nuestra ignorancia de cómo funciona el cerebro. Dado un completo conocimiento del cerebro, me parece probable que pensáramos que es obvio que, si el cerebro estuviera en cierto tipo de estado, debería ser consciente. Obsérvese que ya aceptamos esta forma de necesidad de estados conscientes para ciertos fenómenos de mayor envergadura. Por ejemplo, si veo a un hombre que se queja con el pie atrapado en los dientes de un cepo, sé que debe sufrir un terrible dolor. Es, en cierto sentido, inconcebible que un ser humano normal esté en esa situación sin sufrir un terrible dolor. Las causas físicas hacen necesario el dolor.

Sin embargo, concedámosle a Nagel este extremo, por mor del argumento. Nada se sigue respecto a cómo funciona de hecho el mundo. La limitación que Nagel señala es sólo una limitación de nuestros poderes de concepción. Incluso suponiendo que esté en lo cierto, lo que su argumento muestra es sólo que, en el caso de las relaciones entre cier-

tos fenómenos materiales y otros fenómenos materiales, podemos representarnos subjetivamente los dos términos de la relación; mientras que en el caso de las relaciones entre fenómenos materiales y mentales, un extremo de la relación ya es subjetivo, por lo que no podemos representarnos su relación con lo material del modo en que podemos representarnos la relación entre la liquidez y el movimiento molecular, por ejemplo. El argumento de Nagel, en pocas palabras, sólo muestra que no podemos desembarazarnos de la subjetividad de nuestra conciencia para ver su relación necesaria con su base material. Nos formamos una imagen de la necesidad basada en nuestra subjetividad, pero no podemos formarnos de ese modo una imagen de la necesidad de la relación entre la subjetividad y los fenómenos neurofisiológicos, porque ya estamos en la subjetividad, y la relación imaginativa requeriría que nos saliéramos de ella. (Si la solidez fuera consciente, le parecería misterioso cómo estaría causada por los movimientos vibratorios de las moléculas en ciertas estructuras reticulares y, sin embargo, esos movimientos explican la solidez.)

Podemos apreciar esta objeción a Nagel si imaginamos otras formas de detectar relaciones causales necesarias. Supongamos que Dios o una máquina pudiera detectar simplemente relaciones causales necesarias; en ese caso no existirían, para Dios o la máquina, diferencias entre las formas de necesidad materia-materia y las formas de necesidad materia-mente. Además, incluso aunque concedamos que no podemos representarnos ambos términos de la relación para la conciencia y el cerebro del modo en que podemos representárnoslos para la relación entre liquidez y movimiento molecular, todavía podríamos capturar las relaciones causales involucradas en la producción de la conciencia por medios indirectos. Supongamos que tuviéramos de hecho una descripción de los procesos neurofisiológicos del cerebro que causan la conciencia. No es en modo alguno imposible que lleguemos a tal descripción, porque los procedimientos de prueba usuales para las relaciones causales pueden realizarse respecto a la relación cerebro-conciencia del mismo modo en que pueden realizarse para cualquier fenómeno natural. El conocimiento de las relaciones causales legaliformes nos proporcionará todo lo que necesitemos respecto a la necesidad causal. De hecho, ya tenemos los comienzos de tales relaciones legaliformes. Como mencioné en el capítulo 3, los libros de texto más habituales de neurofisiología explican rutinariamente, por ejemplo, las semejanzas y diferencias entre cómo ven las cosas los gatos y los humanos. No se

cuestiona que ciertos tipos de semejanzas y diferencias neurofisiológicas sean causalmente suficientes para ciertas formas de semejanzas y diferencias en las experiencias visuales. Además, podremos descomponer la gran cuestión —¿Cómo causa la conciencia el cerebro?— en muchas cuestiones más pequeñas (por ejemplo, ¿cómo es que la cocaína produce ciertas experiencias características?). Y las respuestas detalladas que ya estamos comenzando a dar (por ejemplo, la cocaína distorsiona la capacidad de ciertos receptores sinápticos de reabsorber la norepenefrina) permiten las inferencias características asociadas a la necesidad causal (por ejemplo, si se incrementa la dosis de cocaína, se incrementa el efecto). Concluyo que Nagel no ha mostrado que el problema mente-cuerpo sea irresoluble, ni siquiera dentro de nuestro aparato conceptual y nuestra cosmovisión actuales.

Colin McGinn (1991) lleva el argumento de Nagel un paso más lejos y argumenta que es imposible *en principio* que pudiéramos ni siquiera ser capaces de entender la solución al problema mente-cuerpo. Su argumento va más allá del de Nagel e involucra supuestos que Nagel no acepta, al menos no de un modo explícito. Dado que los presupuestos de McGinn son ampliamente compartidos en la tradición filosófica del dualismo, y dado que en este libro trato —entre otras cosas— de superar esos presupuestos, los enunciaré de un modo explícito e intentaré mostrar que son falsos. McGinn supone:

1. La conciencia es un tipo de «material».[6]
2. Este material es conocido por la «facultad de introspección». La conciencia es el «objeto» de la facultad introspectiva, como el mundo físico es el objeto de la facultad perceptiva (pp. 14 y ss., y pp. 61 y ss.)

Es una consecuencia de 1 y 2, aunque no estoy seguro de que McGinn la acepte, que la conciencia como tal, como es conocida por la introspección, no es espacial, en contraste con el mundo físico, que, como tal, como es conocido por la percepción, es espacial.

3. Para poder comprender las relaciones mente-cuerpo, deberíamos comprender el «vínculo» entre la conciencia y el cerebro (*passim*).

6. «Lógicamente, 'conciencia' es un término-de-material, como lo es 'materia'; y no veo nada erróneo, metafísicamente, en el reconocimiento de que la conciencia *es* un tipo de material» (p. 60).

McGinn no duda de que existe tal «vínculo», pero cree que es imposible en principio que lo comprendamos. Dice, usando el término kantiano, que se trata de una relación que, para nosotros, es nouménica. Es imposible para nosotros entender ese vínculo y, por lo tanto, es imposible que entendamos las relaciones mente-cuerpo. McGinn sugiere que el vínculo está proporcionado por una estructura interna de la la conciencia que es inaccesible a la introspección.

Estos tres presupuestos son cartesianos y la «solución» propuesta es una solución de estilo cartesiano (con el inconveniente añadido de que la estructura oculta de la conciencia es, en principio, incognoscible. Al menos la glándula pineal era cognoscible). Sin embargo, como en el caso de la glándula pineal, la solución propuesta no es solución. Si necesitas un vínculo entre la conciencia y el cerebro, necesitas un vínculo entre la estructura oculta de la conciencia y el cerebro. La apelación a la estructura oculta —incluso si fuera inteligible— no nos lleva a ninguna parte.

El problema real está en los tres supuestos. En realidad, creo que involucran la mayoría de los errores del dualismo tradicional durante los últimos trescientos años. Específicamente,

1. La conciencia no es un «material», es un *rasgo* o *propiedad* del cerebro en el sentido, por ejemplo, en que la liquidez es un rasgo del agua.

2. La conciencia no es conocida por introspección de un modo análogo al que los objetos del mundo son conocidos por medio de la percepción. Desarrollo este extremo en el capítulo siguiente y ya lo he comenzado a tratar en éste, así que lo enunciaré de un modo muy simple: el modelo de la inspección interior requiere la distinción entre el acto de inspeccionar y el objeto inspeccionado, y no podemos establecer esa distinción para la conciencia. La doctrina de la introspección es un buen ejemplo de lo que Wittgenstein denomina el embrujamiento de nuestra inteligencia por medio del lenguaje.

Además, una vez que nos hemos desembarazado de la idea de que la conciencia es un material que es «objeto» de la introspección, es fácil ver que es espacial, porque está localizada en el cerebro. En la experiencia consciente no aprehendemos ni la localización espacial ni las dimensiones de nuestra experiencia consciente, pero ¿por qué deberíamos hacerlo? Es una cuestión neurofisiológica extremadamente complicada, que estamos muy lejos de solucionar, la de determinar exacta-

mente cuál es el lugar de la experiencia consciente en nuestro cerebro. Por todo lo que sabemos, podría estar distribuida sobre porciones muy extensas del cerebro.

3. La conciencia y el cerebro no están relacionados por «vínculo» alguno, más de lo que lo están la liquidez del agua y las moléculas de H_2O. Si la conciencia es un rasgo de nivel superior del cerebro, no puede haber cuestión alguna sobre la existencia de un vínculo entre el rasgo y el sistema del que es un rasgo.

IV. LA CONCIENCIA Y LA VENTAJA SELECTIVA

Mi aproximación a la filosofía de la mente, el naturalismo biológico, se enfrenta a veces al reto siguiente. Si se puede imaginar la misma o semejante conducta producida por un zombi inconsciente, ¿por qué la evolución produjo conciencia? De hecho, esto se presenta a menudo como una sugerencia de que es posible que la conciencia ni siquiera exista. Por supuesto, no voy a tratar de demostrar la existencia de la conciencia. Si alguien no está consciente, no hay manera de demostrarle la existencia de la conciencia; si está consciente, es inconcebible que pueda dudar seriamente de que está consciente. No digo que no hay personas que están tan confundidas filosóficamente que *dicen* que dudan de si están conscientes, pero encuentro difícil tomar muy en serio tales enunciados.

Al responder la cuestión relativa al papel de la conciencia en la evolución, deseo rechazar el supuesto implícito de que todo rasgo heredado biológicamente debe proporcionar alguna ventaja evolutiva al organismo. Me parece una forma excesivamente cruda de darwinismo, y tenemos hoy en día todo tipo de buenas razones para abandonarla. Si fuera verdad que toda predisposición innata de un organismo fuera el resultado de alguna presión selectiva, tendría que concluir que mi perro ha sido seleccionado para atrapar pelotas de tenis. Tiene verdadera pasión por atrapar pelotas de tenis y, obviamente, no es algo que haya aprendido, pero esa no es razón para suponer que ello deba tener alguna retribución biológica. O, más próximo a nosotros, creo que la pasión que tienen los seres humanos por esquiar tiene una base biológica que no es el resultado del entrenamiento o del condicionamiento. La extensión del hábito de esquiar ha sido enorme, y los sacrificios que la gente está dispuesta a hacer, en términos de dinero, comodidad y tiempo, a cambio de estar es-

quiando unas pocas horas es, como mínimo, evidencia bastante buena de que obtiene de ello satisfacciones inherentes a su naturaleza biológica. Pero, simplemente, no es el caso que fuéramos seleccionados por la evolución por nuestra predilección para el esquí de montaña.[7]

Con estas matizaciones, todavía podemos enfrentarnos a la cuestión «¿Cuál es la ventaja evolutiva de la conciencia?». Y la respuesta es que la conciencia hace todo tipo de cosas. Para empezar, existen formas de conciencia de todo tipo como la visión, el oído, el gusto, el olfato, la sed, los dolores, los escozores y la acción voluntaria. En segundo lugar, dentro de cada una de estas áreas puede haber una variedad de funciones servidas por las formas conscientes de esas modalidades diferentes. Sin embargo, hablando en los términos más generales, parece claro que la conciencia sirve para organizar cierto conjunto de relaciones entre el organismo, su entorno y sus estados internos. Y, hablando de nuevo en en términos muy generales, la forma de organización podría ser descrita como «representación». Por medio de las modalidades sensoriales, por ejemplo, el organismo obtiene información consciente sobre el estado del mundo. Oye sonidos en su vecindad, ve objetos y estados de cosas en su campo de visión; huele los olores específicos de los distintos rasgos de su entorno; etc. Además de sus experiencias sensoriales conscientes, el organismo también tendrá experiencias características de actuar. Correrá, paseará, comerá, peleará, etc. Esas formas de conciencia no tienen, primordialmente, el propósito de obtener información sobre el mundo; más bien son casos en los que la conciencia capacita al organismo para actuar sobre el mundo, para producir efectos en el mundo. Hablando de nuevo en términos muy generales —y lo discutiremos en términos más refinados más adelante— podemos decir que, en la percepción consciente, el organismo tiene representaciones causadas por los estados de cosas del mundo y, en el caso de la acción intencional, el organismo causa los estados de cosas del mundo por medio de sus representaciones conscientes.

Si esta hipótesis es correcta, podemos hacer una afirmación general sobre la ventaja evolutiva de la conciencia: la conciencia nos proporciona poderes de discriminación mucho mayores de los que tendrían los mecanismos inconscientes de discriminación.

7. La explicación alternativa es que tenemos otros impulsos biológicos más generales que se satisfacen por medio de esas diversas actividades. Compárese la distinción de Elliot Sober (1984, cap. 4) entre lo que se *selecciona* y *aquello para lo que se selecciona.*

Los estudios sobre el caso de Penfield (1975) lo confirman. Algunos de los pacientes de Penfield sufrían una forma de epilepsia conocida como *petit mal* (pequeño mal). En algunos de estos casos, el ataque epiléptico dejaba al paciente totalmente inconsciente, aunque el paciente continuaba exhibiendo lo que normalmente sería considerado como conducta dirigida-a-fines. He aquí algunos ejemplos:

> Un paciente, que denominaré A, era un estudiante serio de piano y estaba sometido a los automatismos del tipo denominado *petit mal.* Era propenso a una interrupción breve mientras practicaba, que era reconocida por su madre como el comienzo de una «ausencia». Luego continuaba interpretando durante un tiempo con considerable agilidad. El paciente B sufría un automatismo epiléptico que comenzaba con una descarga en el lóbulo temporal. A veces el ataque se producía mientras caminaba hacia su casa después de trabajar. Continuaba andando y encontraba su camino de vuelta por entre las calles repletas. Podía darse cuenta más tarde de que había sufrido un ataque porque había una zona en blanco en sus recuerdos correspondientes a parte del trayecto, como desde la calle X a la avenida Y. Si el paciente C estaba conduciendo un coche, continuaba haciéndolo, aunque podía descubrir más tarde que se había saltado algún semáforo en rojo (p. 39).

En todos estos casos, tenemos una forma compleja de conducta aparentemente dirigida a fines sin ninguna conciencia. ¿Por qué no podría ser toda la conducta de este tipo? ¿Qué es lo que añade la conciencia? Advirtamos que, en todos estos casos, los pacientes ejecutaban tipos de acción que eran habituales, rutinarios y memorizados. Es de presumir que existieran caminos neuronales bien establecidos en el cerebro del hombre correspondientes a su conocimiento del camino de vuelta a casa y, de un modo semejante, es de presumir que el conocimiento del pianista de cómo tocar la pieza particular estuviera realizado en los caminos neuronales de su cerebro. La conducta compleja puede ser preprogramada en la estructura del cerebro, al menos en la medida en que sabemos cómo funciona el cerebro en esos casos. Aparentemente, una vez comenzada, la actividad puede seguir su desarrollo incluso bajo el efecto de un ataque de *petit mal.* Pero la conducta consciente humana normal tiene un grado de flexibilidad y creatividad que está ausente en los casos de Penfield del conductor o del pianista inconscientes. La conciencia añade poderes de discriminación y flexibilidad incluso a las actividades rutinarias memorizadas.

Aparentemente, es sólo un hecho biológico el que los organismos que tienen conciencia tienen, en general, poderes de discriminación mayores que aquellos que no la tienen. Los tropismos de las plantas, por ejemplo, que son sensibles a la luz, son mucho menos capaces de hacer discriminaciones refinadas y son mucho menos flexibles que, por ejemplo, el sistema visual humano. La hipótesis que estoy sugiriendo es, pues, la de que una de las ventajas evolutivas que nos confirió el desarrollo de la conciencia es la flexibilidad, la sensibilidad y la creatividad mucho mayores que obtenemos por el hecho de ser conscientes.

Las tradiciones conductistas y mecanicistas que hemos heredado nos ocultan estos hechos; en realidad, hacen que ni siquiera podamos plantear la cuestión de modo apropiado, porque constantemente buscan formas de explicación que tratan lo mental-neurofisiológico como proporcionando simplemente un mecanismo de *input-output*, una función que hace corresponder los estímulos del *input* con las conductas del *output*. Los mismos términos en que se plantean estas cuestiones impiden la introducción de temas que son cruciales para la comprensión de la conciencia, como, por ejemplo, la creatividad.

5. EL REDUCCIONISMO Y LA IRREDUCTIBILIDAD DE LA CONCIENCIA

El punto de vista sobre la relación entre mente y cuerpo que he propuesto es denominado a veces «reduccionismo» y, a veces, «antirreduccionismo». Es denominado «emergentismo» muy a menudo, y es generalmente considerado como una forma de «superveniencia». No estoy seguro de que alguna de esas atribuciones sea en absoluto clara, pero hay un buen número de cuestiones alrededor de estos términos misteriosos y en este capítulo exploraré algunas de ellas.

I. PROPIEDADES EMERGENTES

Supongamos que tenemos un sistema, *S*, compuesto de los elementos *a, b, c*... Por ejemplo, *S* podría ser una piedra y los elementos podrían ser moléculas. En general, habrá rasgos de *S* que no son, o no necesariamente, rasgos de *a, b, c*... Por ejemplo, *S* podría pesar 20 kilogramos, sin que las moléculas individualmente pesen 20 kilogramos. Denominemos a esos rasgos «rasgos del sistema». La forma y el peso de la piedra son rasgos del sistema. Algunos rasgos del sistema pueden ser deducidos, o determinados, o calculados a partir de los rasgos *a, b, c*..., simplemente por la forma en que se componen y ordenan (y, algunas veces, por sus relaciones con el resto del entorno). Ejemplos de ellos serían la forma, el peso y la velocidad. Pero algunos otros rasgos del sistema no pueden ser determinados sólo a partir de los elementos que los componen y de las relaciones con el entorno; han de ser explicados a partir de las relaciones entre los elementos. Llamémosles «rasgos del sistema causalmente emergentes». La solidez, la liquidez y la transparencia son ejemplos de rasgos del sistema causalmente emergentes.

Según estas definiciones, la conciencia es una propiedad causalmente emergente de los sistemas. Es un rasgo emergente de ciertos sistemas de neuronas en el mismo sentido en que la solidez y la liquidez son rasgos emergentes de sistemas moleculares. La existencia de conciencia puede ser explicada por las interacciones causales entre elementos del cerebro al micronivel, pero la conciencia misma no puede ser deducida o calculada a partir de la mera estructura física de las neuronas sin alguna explicación adicional de las relaciones causales entre ellas.

Esta concepción de la emergencia causal, denominémosla «emergente1», ha de distinguirse de una concepción mucho más arriesgada, denominémosla «emergente2». Un rasgo *F* es emergente2 si y sólo si es emergente1 y *F* tiene poderes causales que no pueden ser explicados por las interacciones causales de *a, b, c*... Si la conciencia fuera emergente2, la conciencia podría causar cosas que no podrían ser explicadas por la conducta causal de las neuronas. La idea ingenua es que la conciencia es segregada por la conducta de las neuronas, pero que, una vez segregada, adquiere vida propia.

Debe resultar obvio, a partir del capítulo anterior, que opino que la conciencia es emergente1, pero no emergente2. De hecho, no puedo pensar en nada que sea emergente2, y parece improbable que seamos capaces de encontrar rasgos que sean emergentes2, porque la existencia de tales rasgos violaría incluso el principio más débil de transitividad de la causalidad.

II. Reduccionismo

La mayoría de las discusiones sobre reduccionismo son extremadamente confusas. El reduccionismo como ideal parece haber sido un rasgo de la filosofía positivista de la ciencia, una filosofía que actualmente está desacreditada en muchos aspectos. Sin embargo, todavía sobreviven las discusiones sobre el reduccionismo, y la intuición básica que subyace en el concepto de reduccionismo parece ser la idea de que podría mostrarse que ciertas cosas no podrían ser *nada más que* otras ciertas cosas. El reduccionismo, entonces, lleva a una forma peculiar de relación de identidad que podría denominarse la relación del «nada-más-que»: en general, los *A* pueden reducirse a los *B*, si y sólo si los *A* no son nada más que los *B*.

Sin embargo, incluso dentro de la relación nada-más-que, se quieren decir tantas cosas diferentes por medio de la noción de reducción que es preciso que comencemos haciendo algunas distinciones. Es importante que, desde el principio, tengamos claro cuáles son los *relata* de la relación. ¿Cuál es su supuesto dominio? ¿Objetos, propiedades, teorías, o qué? Encuentro al menos cinco sentidos diferentes de «reducción» —quizás debiera decir cinco formas diferentes de reducción— en la literatura teórica, y quiero mencionar cada uno de ellos para que podamos ver cuáles son relevantes para nuestro análisis del problema mente-cuerpo.

1. *Reducción ontológica*

La forma más importante de reducción es la reducción ontológica. Es la forma en que puede mostrarse que objetos de ciertos tipos no consisten en nada más que objetos de otros tipos. Por ejemplo, se muestra que las sillas no son nada más que colecciones de moléculas. Esta forma es claramente importante en la historia de la ciencia. Por ejemplo, puede mostrarse que los objetos materiales en general no son más que colecciones de moléculas, que los genes no consisten en nada más que moléculas de ADN. Me parece que esta es la forma de reducción a la que apuntan las otras formas.

2. *Reducción ontológica de propiedades*

Esta es una forma de reducción ontológica, pero concierne a propiedades. Por ejemplo, el calor (de un gas) no es nada más que la energía cinética media de los movimientos moleculares. Las reducciones de propiedades para las propiedades correspondientes a los términos teóricos, tales como «calor», «luz», etc., son a menudo el resultado de las reducciones teóricas.

3. *Reducción teórica*

Las reducciones teóricas son las favoritas de los teóricos en sus escritos, pero me parecen más bien extrañas en la práctica real de la cien-

cia, y quizás no sea sorprendente que los manuales más usados repitan hasta la saciedad la misma media docena de ejemplos. Desde el punto de vista de la explicación científica, las reducciones teóricas son enormemente interesantes si permiten llevar a cabo reducciones ontológicas. En cualquier caso, la reducción teórica es primariamente una relación entre teorías, en la que las leyes de la teoría reducida pueden ser (más o menos) deducidas de las leyes de la teoría reductora. Esto demuestra que la teoría reducida no es nada más que un caso especial de la teoría reductora. El ejemplo clásico que suelen ofrecer los manuales es el de la reducción de las leyes de los gases a las leyes de la termodinámica estadística.

4. *Reducción lógica o definicional*

Esta forma de reducción fue durante algún tiempo la más favorecida por los filósofos, pero en los años recientes ha quedado pasada de moda. Se trata de una relación entre palabras y oraciones, donde las palabras y las oraciones que se refieren a un tipo de entidad pueden ser traducidas sin residuo alguno a aquellas que se refieren a otro tipo de entidad. Por ejemplo, las oraciones sobre el fontanero medio de Berkeley son reductibles a oraciones sobre fontaneros individuales específicos de Berkeley; las oraciones sobre números, de acuerdo con una teoría, pueden ser traducidas a, y de ahí reductibles a, oraciones sobre conjuntos. Dado que las palabras y las oraciones son *lógica* o *definicionalmente* reductibles, las entidades correspondientes a las que se refieren las palabras y las oraciones son *ontológicamente* reductibles. Los números, por ejemplo, no son nada más que conjuntos de conjuntos.

5. *Reducción causal*

Se trata de una relación entre cualquier par de tipos de cosas con poderes causales, donde la existencia y, *a fortiori*, los poderes causales de la entidad reducida se muestra que son completamente explicables en términos de los poderes causales de los fenómenos reductores. Así, por ejemplo, algunos objetos son sólidos, y esto tiene consecuencias causales: los objetos sólidos son impenetrables por otros objetos, son resistentes a la presión, etc. Pero esos poderes causales pueden ser cau-

salmente explicados por los poderes causales de los movimientos vibratorios de las moléculas en estructuras reticulares.

Ahora bien, cuando se acusa a los puntos de vista que defiendo de ser reduccionistas —o, a veces, insuficientemente reduccionistas—, ¿cuál de esos diversos sentidos tienen en mente los acusadores? Creo que no se tienen en cuenta ni la reducción teórica ni la lógica. Aparentemente la cuestión es la de si el reduccionismo causal de mi punto de vista lleva —o no lleva— a la reducción ontológica. El punto de vista que mantengo sobre las relaciones mente/cerebro es una forma de reducción causal, tal y como he definido la noción: los rasgos mentales son causados por procesos neurobiológicos. ¿Implica esto la reducción ontológica?

En general, en la historia de la ciencia las reducciones causales que tienen éxito tienden a llevar a reducciones ontológicas. Porque, donde tenemos una reducción causal con éxito, nos limitamos a redefinir la expresión que denota los fenómenos reducidos de tal manera que los fenómenos en cuestión puedan ser identificados con sus causas. Así, por ejemplo, los términos de color se definieron (tácitamente) en términos de la experiencia subjetiva de los perceptores de color; por ejemplo, «rojo» se definió señalando ejemplares ostensivamente, y luego el rojo real se definió como cualquier cosa que pareciera roja a observadores «normales» bajo circunstancias «normales». Pero, una vez que tenemos una reducción causal de los fenómenos de color a la reflexión de la luz, de acuerdo con muchos pensadores, podemos redefinir las expresiones de color en términos de la reflexión luminosa. De este modo, separamos y eliminamos la experiencia subjetiva del color del color «real». El color real ha sufrido una reducción ontológica de propiedades a la reflexión de la luz. Podrían hacerse observaciones similares respecto de la reducción del calor al movimiento molecular, la reducción de la solidez al movimiento molecular en estructuras reticulares, y la reducción del sonido a ondas aéreas. En todos los casos, la reducción causal lleva naturalmente a una reducción ontológica por medio de la redefinición de la expresión que nombra los fenómenos reducidos. Así que, para continuar con el ejemplo de «rojo», una vez que sabemos que las experiencias de color están causadas por cierto tipo de emisión de fotones, tendemos a redefinir la palabra en términos de los rasgos específicos de la emisión de fotones. «Rojo», de acuerdo con algunos autores, se refiere a las emisiones de fotones de 600 nanómetros. Se sigue así, trivialmente, que el color rojo no es más que las emisiones fotónicas de 600 nanómetros.

El principio general en estos casos parece que es el siguiente: una vez que se ve que una propiedad es *emergente1*, tenemos automáticamente una reducción causal que lleva a una reducción ontológica, por redefinición si es preciso. La tendencia general en las reducciones ontológicas que tienen una base científica es hacia una generalidad y objetividad mayores y hacia la redefinición en términos de procesos causales subyacentes.

Todo va bien hasta este punto. Pero nos encontramos ahora con una asimetría que parece sorprendente. Cuando nos las habemos con la conciencia, no podemos realizar la reducción ontológica. La conciencia es una propiedad causalmente emergente de la conducta de las neuronas, y así la conciencia es causalmente reductible a los procesos cerebrales. Pero —y esto es lo que parece tan sorprendente— una ciencia perfecta del cerebro todavía no llevaría a una reducción ontológica de la conciencia del modo en que nuestra ciencia actual puede reducir el calor, la solidez, el color o el sonido. A muchas personas cuya opinión respeto les parece que la irreductibilidad de la conciencia es una razón primordial para que el problema mente-cuerpo continúe siendo tan intratable. Los dualistas tratan la irreductibilidad de la conciencia como prueba incontrovertible de la verdad del dualismo. Los materialistas insisten en que la conciencia debe ser reductible a la realidad material, y en que el precio de negar la reductibilidad de la conciencia sería el abandono de toda nuestra cosmovisión científica.

Expondré brevemente dos cuestiones: en primer lugar, deseo mostrar por qué la conciencia es irreductible, y, en segundo, deseo mostrar por qué el que deba ser irreductible no establece ninguna diferencia en absoluto respecto a nuestra cosmovisión científica. No nos obliga a aceptar el dualismo de propiedades ni nada semejante. Es una consecuencia trivial de ciertos fenómenos más generales.

III. ¿POR QUÉ LA CONCIENCIA ES UN RASGO IRREDUCTIBLE DE LA REALIDAD FÍSICA?

Hay un argumento muy común que trata de mostrar que la conciencia no es reductible del modo en que el calor, etc., lo es. De formas diferentes, el argumento aparece en la obra de Thomas Nagel (1974), Saul Kripke (1971) y Frank Jackson (1982). Pienso que el argumento es decisivo, aunque es mal entendido muchas veces, cuando se le con-

sidera como meramente epistémico y no ontológico. Se le considera a veces como un argumento epistémico que trata de mostrar que, por ejemplo, el tipo de conocimiento de tercera persona, objetivo, que pudiéramos tener de la neurofisiología de un murciélago no incluiría todavía la experiencia subjetiva, de primera persona, de qué se siente al ser un murciélago. Pero para nuestros propósitos actuales, lo fundamental del argumento es ontológico y no epistémico. Trata de los rasgos reales que existen en el mundo y no, excepto en un sentido derivado, de cómo conocemos esos rasgos.

He aquí cómo funciona: consideremos qué hechos del mundo determinan el caso de que tú estés ahora en cierto estado de conciencia como el dolor. ¿Qué hecho del mundo corresponde a tu enunciado verdadero «Ahora tengo dolor»? De un modo ingenuo, parece que hay, como mínimo, dos tipos de hechos. Primero, y más importante, está el hecho de que tú estás teniendo ahora ciertas sensaciones conscientes desagradables, y tú estás experimentando esas sensaciones desde tu punto de vista subjetivo, de primera persona. Son esas sensaciones las que constituyen tu dolor actual. Pero el dolor es también algo causado por ciertos procesos neurofisiológicos subyacentes que consisten en gran parte en patrones de actividad neuronal en el tálamo y otras regiones de tu cerebro. Supongamos que tratamos de reducir la sensación subjetiva consciente, de primera persona, de dolor a los patrones objetivos, de tercera persona, de actividad neuronal. Supongamos que intentáramos decir que el dolor no es «nada más que» los patrones de actividad neuronal. Si intentáramos tal reducción ontológica, los rasgos esenciales del dolor se dejarían de lado. Ninguna descripción de hechos objetivos, fisiológicos, de tercera persona, transmitiría el carácter subjetivo, de primera persona, del dolor, simplemente porque los rasgos de primera persona son diferentes de los rasgos de tercera persona. Nagel establece este extremo al fijar el contraste entre la objetividad de los rasgos de tercera persona y los rasgos de tipo cómo-es de los estados subjetivos de la conciencia. Jackson establece el mismo extremo llamando la atención sobre el hecho de que alguien que tuviera un conocimiento completo de la neurofisiología de fenómenos mentales como el dolor todavía no sabría lo que es el dolor si no supiera cómo-es sentido. Kripke establece el mismo extremo cuando dice que los dolores no podrían ser idénticos a estados neurofisiológicos como la actividad neuronal del tálamo o de cualquier otra parte, porque una identidad tal habría de ser necesaria, porque los dos términos del enunciado de iden-

tidad son designadores rígidos, y, sin embargo, sabemos que la identidad no podría ser necesaria.[1] Este hecho tiene consecuencias epistémicas obvias: mi conocimiento de que tengo dolor tiene una base diferente de la de mi conocimiento de que tú tienes dolor. Pero la consecuencia antirreduccionista más importante es ontológica y no epistémica.

Hasta aquí el argumento antirreduccionista. Es ridículamente simple y definitivo. Se ha vertido una cantidad ingente de tinta tratando de responderlo. Pero las respuestas no son más que tinta desperdiciada. Sin embargo, todavía hay mucha gente a la que le parece que el argumento nos lleva a un callejón sin salida. Les parece que, si aceptamos el argumento, hemos abandonado nuestra cosmovisión científica y hemos aceptado el dualismo de propiedades. En realidad, preguntarían, ¿qué es el dualismo de propiedades más que el punto de vista de que hay propiedades mentales irreductibles? De hecho, ¿no es este el argumento por el que Nagel acepta el dualismo de propiedades y Jackson rechaza el fisicalismo? Y, ¿qué interés tiene el reduccionismo científico si se detiene en las puertas mismas de lo mental? De modo que voy a dedicar mi atención al aspecto crucial de toda esta discusión.

IV. ¿POR QUÉ LA IRREDUCTIBILIDAD DE LA CONCIENCIA NO TIENE CONSECUENCIAS PROFUNDAS?

Para comprender completamente por qué la conciencia es irreductible, debemos considerar con mayor detalle el patrón de reducción que encontramos en las propiedades perceptibles cómo el calor, el sonido, el color, la solidez, la liquidez, etc., y tendremos que mostrar cómo el intento de reducir la conciencia difiere de los otros casos. En cualquier caso, la reducción ontológica estaba basada en una reducción causal previa. Descubríamos que un rasgo superficial de un fenómeno estaba causado por la conducta de los elementos de una microestructura subyacente. Esto es verdad no sólo en los casos en los que el fenómeno reducido era asunto de las apariencias subjetivas, como las «propiedades secundarias» de calor y color, sino también en los casos de «propiedades primarias» como la solidez, en las que había tanto un elemento de apariencia subjetiva (las cosas sólidas son sentidas como sólidas), y

1. Véase el capítulo 2 para una discusión adicional de este extremo.

también muchos rasgos independientes de las apariencias subjetivas (las cosas sólidas, por ejemplo, son resistentes a la presión e impenetrables por otros objetos sólidos). Pero, en cada caso, para las propiedades tanto primarias como secundarias la finalidad de la reducción era aislar los rasgos superficiales y redefinir la noción original en términos de las causas que producen esos rasgos superficiales.

Así, cuando el rasgo superficial es una apariencia subjetiva, redefinimos la noción original de un modo tal que excluimos de su definición las apariencias. Por ejemplo, preteóricamente nuestra noción de calor tiene algo que ver con las temperaturas percibidas. Siendo iguales otros factores, caliente es aquello que es sentido como caliente por nosotros, frío es lo que es sentido como frío. De un modo semejante con los colores: rojo es lo que parece rojo a los observadores normales en circunstancias normales. Pero, cuando tenemos una teoría de lo que causa esos y otros fenómenos, descubrimos que son los movimientos moleculares los que causan las sensaciones de calor y frío (así como otros fenómenos, como incrementos de presión), y la reflexión de la luz la que causa experiencias visuales de cierto tipo (así como otros fenómenos, como los movimientos en los medidores de luz). Entonces, *redefinimos* el calor y el color en términos de las causas subyacentes tanto en las experiencias subjetivas como en otros fenómenos superficiales. Y en la redefinición eliminamos cualquier referencia a las apariencias subjetivas y a otros efectos superficiales de las causas subyacentes. El calor «real» se define ahora en términos de energía cinética de los movimientos moleculares, y la sensación subjetiva de calor que tenemos cuando tocamos un objeto caliente es tratada como la mera apariencia subjetiva causada por el calor, como un efecto del calor. Ya no es parte del calor real. Una distinción similar es la que se hace entre el color real y la apariencia subjetiva de color. El mismo patrón vale para las cualidades primarias: la solidez se define en términos de movimientos vibratorios de moléculas en estructuras reticulares y los rasgos objetivos, independientes del observador, como la impenetrabilidad por otros objetos, son considerados como efectos superficiales de una realidad subyacente. Tales redefiniciones se alcanzan aislando todos los rasgos superficiales del fenómeno, subjetivos u objetivos, y tratándolos como efectos de la cosa real.

Pero, en este punto, debemos advertir lo siguiente: el patrón real de hechos del mundo que corresponde a enunciados sobre formas particulares de calor, como las temperaturas específicas, es bastante semejan-

te al patrón de hechos del mundo que corresponde a los enunciados sobre formas particulares de conciencia, como el dolor. Si digo ahora «Hace calor en esta habitación», ¿cuáles son los hechos? Bueno, en primer lugar, hay un conjunto de hechos «físicos» que involucran los movimientos de las moléculas y, en segundo lugar, hay un conjunto de hechos «mentales» que involucran mi experiencia subjetiva de calor, causada por el impacto de las moléculas del aire en movimiento sobre mi sistema nervioso. Pero sucede algo similar con el dolor. Si digo ahora «Tengo dolor», ¿cuáles son los hechos? Bueno, en primer lugar, hay un conjunto de hechos «físicos» que involucran mi tálamo y otras regiones del cerebro y, en segundo lugar, hay un conjunto de hechos «mentales» que involucran mi experiencia subjetiva de dolor. ¿Por qué consideramos al calor como reductible y al dolor como no reductible? La respuesta es que lo que nos interesa del calor no es la apariencia subjetiva sino las causas físicas subyacentes. Una vez que alcanzamos una reducción causal, redefinimos simplemente la noción para poder llegar a un reducción ontológica. Una vez que conocemos todos los hechos sobre el calor —hechos sobre el movimiento molecular, el impacto sobre terminales nerviosas, las sensaciones subjetivas, etc.—, la reducción del calor a los movimientos moleculares no involucra ningún *hecho* nuevo en absoluto. Es simplemente una consecuencia trivial de la redefinición. No descubrimos primero todos los hechos y después descubrimos un hecho nuevo, el hecho de que el calor es reductible; más bien, redefinimos simplemente el calor de modo que la reducción se siga de la definición. Pero esta redefinición no elimina del mundo, y no se pretende que elimine, las experiencias subjetivas de calor (o color, etc.). Existen como siempre han existido.

Podríamos no haber realizado la redefinición. El obispo Berkeley, por ejemplo, se negó a aceptar tales redefiniciones. Pero es fácil ver por qué es racional hacerlas y aceptar sus consecuencias. Para llegar a una mayor comprensión y un mayor control de la realidad, deseamos saber cómo es el mundo en términos causales, y queremos que nuestros conceptos se adecuen a la naturaleza en sus nexos causales. Redefinimos simplemente los fenómenos con rasgos superficiales en términos de las causas subyacentes. Parece entonces que es un nuevo descubrimiento que el calor no es *nada más que* la energía cinética media del movimiento molecular y que, si desaparecieran del mundo todas las experiencias subjetivas, el calor real todavía permanecería. Pero esto no es un descubrimiento nuevo, es una consecuencia trivial de la nueva defi-

nición. Tales reducciones no muestran que el calor, la solidez, etc., no existen realmente del modo en que, por ejemplo, un conocimiento nuevo mostró que las sirenas y los unicornios no existen.

¿No podríamos decir lo mismo sobre la conciencia? En el caso de la conciencia, no tenemos la distinción entre los procesos «físicos» y las experiencias subjetivas «mentales», ¿por qué la conciencia no puede ser redefinida en términos de procesos neurofisiológicos del modo en que redefiníamos el calor en términos de los procesos físicos subyacentes? Bueno, por supuesto, si insistiéramos en hacer la redefinición podríamos. Podríamos, por ejemplo, definir simplemente «dolor» como patrones de actividad neuronal que causan las sensaciones subjetivas de dolor. Si tal redefinición tuviera lugar, habríamos alcanzado el mismo tipo de reducción para el dolor que tenemos para el calor. Pero, por supuesto, la reducción del dolor a su realidad física todavía deja sin reducir la experiencia subjetiva del dolor. Parte de la finalidad de las reducciones era aislar las experiencias subjetivas y excluirlas de la definición de los fenómenos reales, que se definen en términos de los rasgos que más nos interesan. Pero cuando el fenómeno que más nos interesa son las experiencias subjetivas mismas, no hay manera de aislar nada. Parte de la finalidad de la reducción en el caso del calor era distinguir entre la apariencia subjetiva, por una parte, y la realidad física subyacente, por otra. En realidad, es un rasgo general de tales reducciones que el fenómeno es definido en términos de la «realidad» y no en términos de la «apariencia». Pero no podemos establecer ese tipo de distinción entre apariencia y realidad para la conciencia, porque la conciencia consiste en las apariencias mismas. *Cuando se trata de la apariencia no podemos realizar la distinción entre apariencia y realidad porque la apariencia es la realidad.*

Para el propósito que nos ocupa, podemos resumir este extremo diciendo que la conciencia no es reductible del modo en que otros fenómenos son reductibles, no porque el patrón de hechos en el mundo real involucre nada especial, sino porque la reducción de otros fenómenos dependía, en parte, de la distinción entre «realidad física objetiva», por un lado, y la mera «apariencia subjetiva», por otro, y de eliminar la apariencia en los fenómenos que han sido reducidos. Pero, en el caso de la conciencia, su realidad es la apariencia; de modo que la finalidad de la reducción se perdería si intentáramos aislar la apariencia y definiéramos simplemente la conciencia en términos de la realidad física subyacente. En general, el patrón de nuestras reducciones descansa en el recha-

zo de la base subjetiva epistémica a la presencia de una propiedad como parte del último constituyente de esa propiedad. Descubrimos cosas sobre el calor o la luz sintiendo y viendo, pero definimos luego el fenómeno de un modo que es independiente de la epistemología. La conciencia es una excepción a este patrón por una razón trivial. La razón, para repetirlo, es que las reducciones que dejan a un lado las bases epistémicas, las apariencias, no pueden funcionar para las bases epistémicas mismas. En tales casos, la apariencia es la realidad.

Pero esto muestra que la irreductibilidad de la conciencia es una consecuencia trivial de la pragmática de nuestras prácticas definicionales. Un resultado trivial como este sólo tiene consecuencias triviales. No tiene consecuencias metafísicas profundas para la unidad de nuestra cosmovisión científica global. No muestra que la conciencia no sea parte del mobiliario último de la realidad, o que no pueda ser objeto de la investigación científica, o que no pueda ser incorporada a nuestra concepción física global del universo; muestra meramente que la conciencia, por definición, está excluida de cierto patrón de reducción, dada la manera en la que hemos decidido llevar a cabo la reducción. La conciencia no puede ser reductible, no a causa de ningún rasgo misterioso, sino simplemente porque, por definición, cae fuera del patrón de reducción que hemos escogido por razones pragmáticas. Preteóricamente, la conciencia, como la solidez, es un rasgo superficial de ciertos sistemas físicos. Pero, a diferencia de la solidez, la conciencia no puede ser redefinida en términos de una microestructura subyacente, y los rasgos superficiales tratados después como meros efectos de la conciencia real, sin perder el motivo por el que tenemos el mismo concepto de conciencia.

Hasta ahora, el argumento de este capítulo ha sido desarrollado, por decirlo de algún modo, desde el punto de vista del materialista. Podemos resumir mis observaciones del siguiente modo: el contraste entre la reductibilidad del calor, el color, la solidez, etc., por una parte, y la irreductibilidad de los estados conscientes, por otra, no refleja ninguna distinción en la estructura de la realidad, sino una distinción en nuestras prácticas definicionales. Podríamos expresar el mismo extremo desde el punto de vista del dualismo de propiedades del modo siguiente: el contraste aparente entre la irreductibilidad de la conciencia y la reductibilidad del color, el calor, la solidez, etc., en realidad era sólo *aparente*. No eliminábamos realmente la subjetividad del rojo, por ejemplo, cuando reducíamos el rojo a las reflexiones de la luz; simplemente de-

jábamos de denominar «rojo» a ese componente subjetivo. No eliminábamos ningún fenómeno subjetivo con esas «reducciones», simplemente dejábamos de denominarlo por medio de los viejos nombres. Tanto si tratamos la irreductibilidad desde el punto de vista del materialista como del dualista, todavía se nos deja con un universo que contiene un componente físico irreductiblemente subjetivo como componente de la realidad física.

Para concluir esta parte de la discusión, deseo que quede claro tanto lo que estoy diciendo como lo que no estoy diciendo. No estoy diciendo que la conciencia no es un fenómeno extraño y maravilloso. Creo, por el contrario, que debemos sentirnos perplejos por el hecho de que los procesos evolutivos produjeran sistemas nerviosos capaces de causar y mantener estados conscientes subjetivos. Como señalé en el capítulo 4, la conciencia es tan empíricamente misteriosa para nosotros como lo fue anteriormente el electromagnetismo, cuando la gente pensó que el universo debía operar completamente de acuerdo con los principios newtonianos. Pero estoy diciendo que, una vez que se admite la existencia de conciencia (subjetiva, cualitativa) —y nadie en su sano juicio puede negarla, aunque algunos finjan hacer tal cosa—, no hay nada extraño, maravilloso o misterioso sobre su *irreductibilidad*. Su irreductibilidad no tiene en absoluto desafortunadas consecuencias para la ciencia. Además, cuando hablo de la irreductibilidad de la conciencia, estoy hablando de su *irreductibilidad de acuerdo con patrones comunes de reducción*. Nadie puede eliminar *a priori* la posibilidad de una enorme revolución intelectual que nos diera una concepción nueva —por el momento inimaginable— de la reducción, de acuerdo con la cual la conciencia fuera reductible.

V. Superveniencia

En los últimos años, se ha discutido mucho una relación entre propiedades denominada «superveniencia» (por ejemplo, Kim, 1979, 1982; Haugeland, 1982). Se dice a menudo en las discusiones de filosofía de la mente que lo mental sobreviene a lo físico. Intuitivamente, lo que se quiere decir es que los estados mentales son completamente dependientes de los estados neurofisiológicos correspondientes, en el sentido de que una diferencia en los estados mentales involucraría necesariamente una diferencia correspondiente en los estados neurofisiológicos.

Si por ejemplo, paso de un estado de tener sed a un estado de no tener sed, debe haber algún cambio en mis estados cerebrales correspondiente al cambio en mis estados mentales.

Según el análisis que he venido defendiendo, los estados mentales sobrevienen a los estados neurofisiológicos en el siguiente sentido: causas neurofisiológicas idénticas en tipo tendrían efectos mentales idénticos en tipo. Así, por considerar el famoso ejemplo del cerebro-en-una-cubeta, si tuviéramos dos cerebros que fueran idénticos en tipo hasta la última molécula, las bases causales de lo mental garantizarían que tuvieran los mismos fenómenos mentales. Según esta caracterización de la relación de superveniencia, la superveniencia de lo mental sobre lo físico está marcada por el hecho de que los estados físicos son causalmente suficientes, aunque no necesarios, para los estados mentales correspondientes. Se trata de otra manera de decir que, en la medida en que está implicada esta definición de superveniencia, la identidad en la neurofisiología garantiza la identidad en lo mental; pero que la identidad en lo mental no garantiza la identidad en lo neurofisiológico.

Vale la pena poner de relieve que esta clase de superveniencia es superveniencia *causal*. Las discusiones sobre la superveniencia se introdujeron en un principio en conexión con la Ética, y la noción involucrada no era una noción causal. En los primeros escritos de Moore (1922) y Hare (1952), la idea fue la de que las propiedades morales sobrevienen a las propiedades naturales, que dos objetos no pueden diferir solamente, por ejemplo, en su bondad. Si un objeto es mejor que otro, debe existir algún otro rasgo en virtud del cual el primero es mejor que el segundo. Pero esta noción de superveniencia moral no es una noción causal. Esto es, los rasgos de un objeto que lo hacen bueno no *causan* que sea bueno, más bien *constituyen* su bondad. Pero en el caso de la superveniencia mente/cuerpo, los fenómenos neuronales causan los fenómenos mentales.

De modo que hay, al menos, dos nociones de superveniencia: una noción constitutiva y una noción causal. Creo que sólo la noción causal es importante para las discusiones del problema mente-cuerpo. En este respecto, mi análisis difiere de los análisis más habituales de la superveniencia de lo mental sobre lo físico. Así, Kim (1979, especialmente pp. 45 y ss.) pretende que no deberíamos pensar en la relación de los sucesos neuronales con sus sucesos mentales sobrevenidos como una relación causal y, de hecho, pretende que los sucesos mentales sobrevenidos no tienen un estatus causal además de su superveniencia res-

pecto a sucesos neurofisiológicos que tienen «un papel causal más directo». «Si esto es epifenomenalismo, saquemos de él lo que podamos», dice tranquilamente (p. 47).

Estoy en desacuerdo con estas dos afirmaciones. Me parece obvio, por todo lo que sabemos sobre el cerebro, que todos los fenómenos macromentales están causados por microfenómenos de nivel inferior. No hay nada misterioso en esa relación causal de abajo-arriba; es muy común en el mundo físico. Además, el hecho de que los rasgos mentales sobrevengan a los rasgos neuronales no disminuye en modo alguno su eficacia causal. La solidez del pistón sobreviene causalmente a su estructura molecular, pero esto no hace que la solidez sea epifenoménica; y, de un modo semejante, la superveniencia causal de mi dolor de espalda sobre microsucesos cerebrales no hace que el dolor sea epifenómenico.

Mi conclusión es que, una vez que reconocemos la existencia de formas de causalidad de abajo-arriba, micro-macro, la noción de superveniencia no juega ningún papel importante en filosofía. Los rasgos formales de la relación ya están presentes en la suficiencia causal de las formas de causalidad micro-macro. Y la analogía con la Ética es sólo una fuente de confusión. La relación de los rasgos macromentales del cerebro con sus rasgos microneuronales es totalmente diferente de la relación de la bondad con los rasgos que determinan la bondad, y es confuso colocarlas en el mismo cajón. Como dice Wittgenstein en algún lugar, «Si envuelves tipos de muebles diferentes con suficiente papel, puedes conseguir que parezca que todos tienen la misma forma».

6. LA ESTRUCTURA DE LA CONCIENCIA: UNA INTRODUCCIÓN

He hecho de pasada algunas afirmaciones sobre la naturaleza de la conciencia, y es tiempo ahora de intentar una explicación más general. Tal tarea puede parecer a la vez imposiblemente difícil y ridículamente fácil. Difícil puesto que, despues de todo, ¿no es la historia de nuestra conciencia la historia de toda nuestra vida? Y fácil porque, después de todo, ¿no estamos más próximos a la conciencia que a ninguna otra cosa? De acuerdo con la tradición cartesiana, tenemos conocimiento cierto e inmediato de nuestros estados conscientes, de modo que la tarea debería ser fácil. Pero no lo es. Por ejemplo, encuentro que es fácil describir los objetos de la mesa que hay enfrente de mí, ¿pero cómo describiría separadamente y de manera adicional mi experiencia consciente de esos objetos?

Hay dos temas que son cruciales para la conciencia, pero tendré poco que decir sobre ellos puesto que todavía no los entiendo bien. El primero es la temporalidad. Desde Kant hemos sido conscientes de una asimetría en la manera en que la conciencia se relaciona con el espacio y el tiempo. Aunque experimentamos los objetos y los eventos como algo que es, a la vez, espacialmente extendido y de duración temporal, nuestra conciencia misma no se experimenta como algo espacial, aunque se experimenta como algo que es temporalmente extendido. De hecho, las metáforas espaciales para describir el tiempo parecen casi también inevitables para la conciencia, como cuando hablamos, por ejemplo, del «flujo de la conciencia». Notoriamente, el tiempo fenomenológico no encaja exactamente con el tiempo real, pero no sé cómo dar cuenta del carácter sistemático de las disparidades.[1]

1. Incluso asuntos obvios como que, cuando uno está aburrido, «el tiempo pasa más despacio» me parece que exigen alguna explicación. ¿Por qué el tiempo tiene que pasar más despacio cuando uno está aburrido?

El segundo asunto que se olvida es la sociedad. Estoy convencido de que la categoría de «otras personas» juega un papel especial en la *estructura* de nuestras experiencias conscientes, un papel distinto del de los objetos y estados de cosas; y creo que esta capacidad para asignar un estatus especial a otros *loci* de la conciencia está basada biológicamente y es, a la vez, una presuposición de Trasfondo para todas las formas de intencionalidad colectiva (Searle, 1990). Pero todavía no sé cómo demostrar estas afirmaciones ni cómo analizar la estructura de los elementos sociales en la conciencia individual.

I. Una docena de rasgos estructurales

En lo que sigue, intentaré describir rasgos estructurales gruesos de la conciencia normal, cotidiana. A menudo, el argumento que usaré para identificar un rasgo es la ausencia del rasgo en formas patológicas.

1. *Modalidades finitas*

La conciencia humana se manifiesta en un número estrictamente limitado de modalidades. Además de los cinco sentidos de la vista, el tacto, el olfato, el gusto y el oído, y el sexto, el «sentido del equilibrio», hay también sensaciones corporales (la «percepción propia») y el flujo del pensamiento. Por sensaciones corporales no entiendo sólo sensaciones físicas tales como los dolores, sino también el que me dé cuenta a traves de los sentidos de, por ejemplo, la posición de mis brazos y piernas o el que sienta algo en mi rodilla derecha. El flujo del pensamiento no sólo contiene palabras e imágenes, sino también otros elementos tanto visuales como de otro tipo, que no son ni verbales ni tienen forma de imágenes. Por ejemplo, un pensamiento le ocurre a uno de repente algunas veces, «se le enciende a uno la bombilla», de una forma tal que no está ni en palabras ni en imágenes. Además, el flujo del pensamiento, tal como estoy usando aquí esta expresión, incluye sentimientos tales como aquellos que se denominan, generalmente, «emociones». Por ejemplo, podría sentir en el flujo del pensamiento una súbita oleada de rabia, o un deseo de golpear a alguien o un fuerte deseo de un vaso de agua.

No hay ninguna razón *a priori* por la que la conciencia deba limi-

tarse a estas formas. Parece ser un hecho de la historia evolutiva humana que estas son las formas que ha desarrollado nuestra especie. Hay una buena evidencia de que otras determinadas especies han desarrollado otras modalidades sensoriales. La visión es especialmente importante en los seres humanos, y de acuerdo con algunas explicaciones neurofisiológicas, más de la mitad de nuestro córtex está dedicado a funciones visuales.

Otro rasgo general de cada modalidad es que puede ocurrir bajo el aspecto de agradable o desagradable, y el modo en que es agradable/desagradable es, en general, algo específico de la modalidad. Por ejemplo, los olores agradables no lo son de la manera en que los pensamientos agradables lo son, incluso si son pensamientos agradables sobre olores agradables. A menudo, pero no siempre, el aspecto placer/displacer de las modalidades conscientes se asocia con una forma de intencionalidad. Así, en el caso de las experiencias visuales, lo que es agradable o desagradable es, en general, la intencionalidad externa a las experiencias visuales más bien que sus aspectos sensoriales puros. Encontramos desagradable ver algo nauseabundo como, por ejemplo, un hombre vomitando; y encontramos agradable ver algo impresionante como, por ejemplo, las estrellas en una noche clara. Pero en cada caso es mucho más que los aspectos puramente visuales de la escena lo que constituye la fuente del carácter agradable o desagradable. Este no es siempre el caso con las sensaciones corporales. El dolor puede experimentarse simplemente como doloroso, sin ninguna intencionalidad que esté correlacionada con él. Sin embargo, el que el dolor no sea agradable es algo que variará con ciertas clases de intencionalidad asociada. Si uno cree que el dolor está siendo causado injustamente es más desagradable que si uno cree que está siendo infligido, por ejemplo, como parte de un tratamiento médico que es necesario. Los orgasmos son algo que se colorea de manera similar por la intencionalidad. Podría imaginarse fácilmente un orgasmo que ocurre sin pensamiento erótico alguno —supóngase, por ejemplo, que ha sido inducido por medios eléctricos—, pero en general el placer de un orgasmo se relaciona internamente con su intencionalidad, aunque los orgasmos sean sensaciones corporales. En esta sección me intereso solamente por el placer/displacer de cada modalidad. Discutiré como rasgo 12 el placer/displacer de los estados conscientes totales.

2. *Unidad*

Es característico de los estados conscientes no patológicos el que nos acaezcan como parte de una secuencia unificada. No tengo sólo una experiencia de un dolor de muelas y también una experiencia visual del sofá que está a unos pocos centímetros de mí y de las rosas que asoman por encima del jarrón que está a mi izquierda, al modo en que llevo puesta una camisa de rayas al mismo tiempo que unos calcetines de color azul oscuro. La diferencia crucial es esta: tengo mis experiencias de la rosa, el sofá y el dolor de muelas como experiencias que son, todas ellas, parte de uno y el mismo evento consciente. La unidad existe en, al menos, dos dimensiones que, continuando con las metáforas espaciales, llamaré «horizontal» y «vertical». La unidad horizontal es la organización de las experiencias conscientes a través de tramos de tiempo cortos. Por ejemplo. cuando digo una oración o la pienso, incluso en el caso de una oración larga, m darme cuenta del comienzo de lo que he dicho o pensado continúa incluso cuando esa parte ya no está siendo pensada o dicha. La memoria icónica de esta clase es esencial para la unidad de la conciencia, y quizás incluso es esencial la memoria a corto plazo. La memoria vertical es u asunto que tiene que ver con el darse cuenta simultáneamente de todos lo diversos rasgos de cualquier estado consciente, tal como ilustra mi ejemplo del sofá, el dolor de muelas y la rosa. Tenemos poca comprensión d cómo el cerebro logra esta unidad. En neurofisiología se denomina a est «el problema del vínculo» («*the binding problem*») y Kant llamó al mismo fenómeno «la unidad trascendental de apercepción».

Sin estos dos rasgos —la unidad horizontal del presente recordado y la unidad vertical de la vinculación de los elementos en una column unificada— no podríamos dar un sentido normal a nuestras experiencias. Esto resulta ilustrado por varias formas de patología tal como lo fenómenos del cerebro partido (*the split-brain*) (Gazzaniga, 1970) y e síndrome de Korsakov (Sacks, 1985).

3. *Intencionalidad*

La mayor parte de la conciencia, si no toda, es intencional. Puedo, po ejemplo, estar simplemente con el ánimo deprimido o alegre sin estar de

2. Esta expresión se debe a Edelman (1991).

primido o alegre sobre nada en particular. En estos casos, mi estado de ánimo, en cuanto que tal, no es intencional. Pero en general, en cualquier estado consciente, el estado se dirige hacia una u otra cosa, incluso si aquello a lo que se dirige no existe, y en este sentido tiene intencionalidad. Para una clase de casos muy extensa, la conciencia es, efectivamente, conciencia de algo y el «de» en «conciencia de» es el «de» de intencionalidad.

La razón por la que encontramos que es difícil distinguir entre mi descripción de los objetos que hay encima de la mesa y mi descripción de mi experiencia de los objetos es que los rasgos de los objetos son precisamente las condiciones de satisfacción de mis experiencias conscientes de ellos. Así, el vocabulario que uso para describir la mesa —«Hay una lámpara a la derecha, un jarrón a la izquierda y una estatuilla en el centro»— es precisamente aquello que uso para describir mis experiencias visuales conscientes de la mesa. Para describir las experiencias tengo que decir, por ejemplo, «Me parece visualmente que hay una lámpara a la derecha, un jarrón a la izquierda y una estatuilla en el centro».

Mis experiencias conscientes, a diferencia de los objetos de las experiencias, tienen siempre una perspectiva. Son siempre experiencias conscientes desde un punto de vista. Perspectiva y punto de vista son obvios sobre todo en el caso de la visión, pero son también, desde luego, rasgos de otras experiencias sensoriales nuestras. Si toco la mesa, tengo una experiencia de ella sólo bajo ciertos aspectos y desde una cierta localización espacial. Si oigo un sonido, lo oigo desde una cierta dirección y oigo ciertos aspectos suyos. Y así sucesivamente.

Obsérvese que el carácter de la experiencia consciente consistente en tener perspectiva es una buena manera de recordarnos que *toda intencionalidad tiene un aspecto.* Ver un objeto desde, por ejemplo, un punto de vista es verlo bajo ciertos aspectos y no bajo otros. En este sentido, todo ver es «ver como». Y lo que vale para ver vale para todas las formas de intencionalidad, conscientes e inconscientes. Todas las representaciones representan sus objetos, u otras condiciones de satisfacción, bajo aspectos. Todo estado intencional tiene lo que yo llamo *un contorno de aspecto.*

4. *Sentimiento subjetivo*

La discusión de la intencionalidad lleva de manera natural al sentimiento subjetivo de nuestros estados conscientes. En capítulos anterio-

res, he tenido la ocasión de discutir la subjetividad con algún detalle, de modo que no volveré a elaborar lo que ya he dicho. Baste decir aquí que la subjetividad involucra necesariamente el aspecto a-qué-se-parece-lo-que-se-siente de los estados conscientes. Así, por ejemplo, puedo razonablemente preguntar a qué se parece el sentirse un delfín y estar nadando todo el día, porque suponemos que los delfines tienen experiencias conscientes. Pero no tiene sentido preguntarse a qué se parece sentirse una ripia clavada al techo año tras año, porque en el sentido en que usamos esta expresión no hay nada a lo que se parezca sentirse una ripia dado que las ripias no son conscientes.

Como he señalado antes, la subjetividad es responsable, más que cualquier otra cosa, del problema filosófico que tiene que ver con la conciencia.

5. *La conexión entre conciencia e intencionalidad*

Espero que la mayor parte de lo que he dicho hasta ahora parezca obvio. Quiero hacer ahora una afirmación radical que no voy a substanciar completamente hasta el próximo capítulo. La afirmación es esta: sólo un ser que pueda tener estados intencionales conscientes puede tener estados intencionales, y todo estado intencional inconsciente es, al menos, potencialmente consciente. Esta tesis tiene enormes consecuencias para el estudio de la mente. Implica, por ejemplo, que cualquier análisis de la intencionalidad que deje fuera la cuestión de la conciencia tendrá que ser incompleta. Es posible describir la estructura lógica de los fenómenos intencionales sin tratar la conciencia —en realidad lo hice así en gran parte en *Intencionalidad* (Searle, 1983)—, pero existe una conexión conceptual entre conciencia e intencionalidad que tiene la consecuencia de que una teoría completa de la intencionalidad exige una explicación de la conciencia.

6. *La base figurativa, la estructura gestáltica de la conciencia*

Desde el desarrollo de la psicología de la *Gestalt* es un asunto que se considera familiar el que nuestras experiencias perceptivas nos llegan como una figura sobre un trasfondo. Por ejemplo, si veo el jersey que está encima de la mesa que hay delante de mí, veo el jersey sobre

el trasfondo de la mesa. Si veo la mesa, la veo sobre el trasfondo del suelo. Y si veo el suelo, lo veo sobre el trasfondo de toda la habitación; y así hasta que alcanzamos los límites de mi campo visual. Pero lo que es característico de la percepción parece ser característico de la conciencia en general: que cualquier cosa en la que concentro mi atención está sobre un trasfondo que no es el centro de atención; y que cuanto más amplio es el alcance de la atención, más cerca estamos de alcanzar los límites de mi conciencia donde el trasfondo serán simplemente las condiciones límite que trataré más adelante como rasgo número 10.

El hecho de que nuestras percepciones normales estén siempre estructuradas está relacionado con la estructura de base figurativa de las experiencias conscientes; el que yo no perciba pura y simplemente contornos indiferenciados, sino que mis percepciones estén organizadas en objetos y rasgos de objetos. Esto tiene como consecuencia el que todo ver (normal) es *ver como*, todo percibir (normal) es *percibir como*, y de hecho, toda conciencia es *conciencia de algo como tal y tal.*

Hay aquí dos rasgos diferentes aunque relacionados. Uno es, dicho de manera general, la estructura de base figurativa de la percepción, y el segundo es la organización de nuestras experiencias perceptivas (y de otras clases) conscientes. La estructura de base figurativa es un caso especial, aunque muy extendido, del rasgo más general de la estructuración. Otro rasgo relacionado, que discutiré brevemente como rasgo número 10, lo constituyen las condiciones límite generales que parecen ser aplicables a todo estado consciente.

7. *El aspecto de la familiaridad*

Dada la temporalidad, carácter social, unidad, intencionalidad, subjetividad y estructuración de la conciencia, me parece que uno de los rasgos más extendidos de los estados ordinarios y no patológicos en los que alguien se da cuenta conscientemente de algo es lo que voy a llamar «el aspecto de la familiaridad». Como toda intencionalidad consciente tiene un aspecto (rasgo 3), y puesto que las formas no patológicas de la conciencia están estructuradas u organizadas (rasgo 6), la posesión previa de un aparato suficiente para generar conciencia organizada y dotada de aspecto garantiza automáticamente que los rasgos de aspecto de la experiencia consciente y las estructuras y la organiza-

ción de la conciencia serán más o menos familiares, de maneras que intentaré explicar ahora.

Podemos captar mejor el aspecto de la familiaridad contrastando mi explicación con la de Wittgenstein. Wittgenstein nos pregunta (1953) si cuando entro en mi habitación experimento un «acto de reconocimiento», y nos recuerda que, de hecho, no hay tal acto. Creo que en esto tiene razón. Sin embargo, no cabe duda de que cuando entro en mi habitación ésta *me parece familiar*. Esto puede verse si nos imaginamos que hubiese algo radicalmente no familiar, si hubiese, por ejemplo, un enorme elefante en medio de la habitación, si el techo se hubiese desplomado o si alguien hubiera cambiado completamente los muebles. Pero en el caso normal, la habitación me parece familiar. Ahora bien, lo que es verdad de mi experiencia de la habitación es verdad, sugiero, en mayor o menor grado de mis experiencias del mundo. Cuando paseo por la calle esos objetos me son familiares como casas, y esos otros objetos me son familiares como gente. Experimento como parte de lo familiar los árboles, la acera, las calles. E incluso cuando estoy en una ciudad extraña y me sorprende la rareza de los vestidos de sus habitantes o lo singular de la arquitectura de sus casas, ahí está, con todo, el aspecto de la familiaridad. Eso es todavía gente; aquellas son todavía casas; yo soy todavía un ser corporal, con un sentido consciente de mi propio peso, un sentido de las fuerzas de gravedad que actúan sobre mí y sobre otros objetos; tengo un sentido interno de mis partes corporales y de sus posiciones. Y quizás lo más importante de todo, tengo un sentido interno de aquello a lo que se parece el que me sienta yo, un sentimiento de mí mismo.[3]

Se requiere cierto esfuerzo intelectual para romper este aspecto de la familiaridad. Así, por ejemplo, los pintores surrealistas pintan paisajes en los que no hay objetos familiares. Pero incluso en tales casos, aún sentimos objetos en un entorno, un horizonte de la tierra, la atracción gravitatorial de los objetos hacia la tierra, la luz que viene de una fuente, un punto de vista desde el que se pinta el cuadro, nos sentimos a no-

3. Hume, dicho sea de paso, pensó que no podía haber tal sentimiento, puesto que si lo hubiese, éste tendría que llevar a cabo una gran cantidad de trabajo epistémico y metafísico que el mero sentimiento no podría llevar a cabo. Pienso que todos nosotros tenemos de hecho un sentido característico de nuestra propia personalidad, aunque tiene muy poco interés epistémico o metafísico. No garantiza la «identidad personal», «la unidad del yo», o cualquier otra cosa por el estilo. Es justamente cómo, por ejemplo, siento yo que soy yo.

sotros mismos viendo el cuadro, etc.; y todo ese sentir es parte del aspecto de la familiaridad de nuestra conciencia. El reloj fláccido es aún un reloj, la mujer de tres cabezas es aún una mujer. Es este aspecto de la familiaridad —más que, por ejemplo, la predictibilidad inductiva— la que impide que los estados conscientes sean la «condenada y zumbona confusión» descrita por William James.

He estado usando deliberadamente la expresión «aspecto de la familiaridad» más bien que la más coloquial «sentimiento de familiaridad» porque quiero señalar que el fenómeno que estoy analizando no es un sentimiento separado. Cuando, por ejemplo, veo mis zapatos no tengo una experiencia visual de los zapatos y, a la vez, un sentimiento de familiaridad, sino más bien lo que sucede es que *veo* los zapatos *como* zapatos y, a la vez, *como* míos. El aspecto de la familiaridad no es una experiencia separada y este es el motivo por el que Wittgenstein tiene razón al decir que no hay un acto de reconocimiento cuando veo mi habitación. Sin embargo, a mí me parece que es como mi habitación, y la percibo bajo este aspecto de familiaridad.

El aspecto de la familiaridad tiene grados diversos; es un fenómeno escalonado. En la parte superior de la escala de familiaridad están los objetos, los escenarios, la gente y las visiones de mi vida ordinaria, cotidiana. Más abajo están las escenas extrañas en las que los objetos y la gente me son, con todo, fácilmente reconocibles y categorizables. Más abajo aún están las escenas en las que encuentro poco que sea reconocible o categorizable. Estas son las clases de escenas pintadas por los pintores surrealistas. Es posible imaginar un caso límite en el que no se perciba nada como familiar, en el que no haya nada reconocible y categorizable, ni siquiera como objeto, donde incluso mi propio cuerpo ya no fuera categorizable como mío y ni siquiera como un cuerpo. Tal caso sería patológico en extremo. Ocurren formas menos extremas de patología cuando, por ejemplo, en los estados de desesperación neurótica uno se fija en la textura de la madera de una mesa y se encuentra totalmente perdido en ella, como si nunca hubiera visto antes semejante cosa.

El aspecto de la familiaridad es lo que hace posible gran parte de la organización y el orden de mis experiencias conscientes. Incluso si encuentro un elefante en mi habitación o me topo con el techo derrumbado, el objeto me es todavía familiar como elefante o como techo derrumbado y la habitación como mi habitación. Los psicólogos tienen una gran cantidad de evidencia para mostrar que la percepción es una

función de expectativas (por ejemplo, Postman, Bruner y Walk, 1951). Un corolario natural de esta afirmación es que la organización de la percepción es sólo posible dado un conjunto de categorías que identifican entidades con lo familiar.

Pienso que el rasgo de la experiencia al que estoy aludiendo será reconocible por cualquiera que piense sobre él, pero describir la estructura de la intencionalidad que está involucrada aquí es bastante más complicado. Los objetos y estados de cosas son experimentados por mí como familiares, pero la familiaridad no es, en general, una condición de satisfacción separada. Más bien, la conciencia incluye categorización —veo cosas, por ejemplo, como árboles, gente, casas, coches, etc.—, pero las categorías tienen que existir antes de la experiencia, puesto que son las condiciones de posibilidad para tener precisamente esas experiencias. Para ver esto como un pato o como un conejo, tengo que tener las categorías «pato» o «conejo» antes de la percepción. Así pues, la percepción procederá bajo el aspecto de la familiaridad, puesto que las categorías que la hacen posible son en sí mismas categorías familiares. El argumento es, en esencia, este: todo percibir es percibir como, y más generalmente, toda conciencia *de* es conciencia *como.* Para ser consciente de algo se tiene que ser consciente de ello como algo (eliminando de nuevo la patología y cosas similares), pero percibir como, y otras formas de conciencia como, exigen categorías. Pero las categorías preexistentes implican familiaridad anterior con las categorías y, por lo tanto, las percepciones lo son bajo el aspecto de lo familiar. *Así pues, los siguientes rasgos están conectados entre sí: estructuración, percepción como, el contorno de aspecto de toda la intencionalidad, y el aspecto de la familiaridad. Las experiencias conscientes nos vienen estructuradas, y esas estructuras nos capacitan para percibir cosas bajo aspectos, pero esos aspectos están constreñidos por nuestro dominio de un conjunto de categorías, y esas categorías, que nos son familiares, nos capacitan, en grados diversos, para asimilar nuestras experiencias, por nuevas que sean, a lo familiar.*

No estoy presentando aquí el argumento falaz de que puesto que tenemos experiencias bajo aspectos familiares tenemos, por tanto, experiencias del aspecto de la familiaridad. Esto no es en absoluto lo que se está discutiendo. Lo que está en juego es más bien que las formas no patológicas de la conciencia tienen, de hecho, un aspecto de la familiaridad, y de esto da cuenta el hecho de que tenemos capacidades de Trasfondo, neurobiológicamente realizadas, para generar experiencias

que sean a la vez estructuradas y con un aspecto, en las que las estructuras específicas y los aspectos sean más o menos familiares. Las capacidades en cuestión no son parte de la conciencia sino del Trasfondo (sobre el Trasfondo véase el capítulo 8).

8. *Desbordamiento*

Los estados conscientes se refieren, en general, a algo que está más allá de su contenido inmediato. Llamo a este fenómeno «desbordamiento». Considérese una clase extrema de caso. Sally mira a Sam y, de repente, tiene un pensamiento instantáneo: «Ya está». Si se le pidiese que enunciase el pensamiento, podría comenzar diciendo: «Bien, me di cuenta de repente de que durante los últimos dieciocho meses había estado perdiendo el tiempo en una relación con alguien que es totalmente inapropiado para mí, que, cualesquiera que sean sus otros méritos, mi relación con Sam estaba basada en una falsa premisa por mi parte. De repente me di cuenta de que no podría tener jamás una relación estable con el jefe de una cuadrilla de moteros como los Ángeles del Infierno porque...». Y así sucesivamente.

En tal caso, el contenido inmediato tiende a rebosar, a conectar con otros pensamientos que en algún sentido eran parte del contenido pero en otro no lo eran. Aunque se ilustra mejor con un caso extremo como este, pienso que el fenómeno es general. Si, por ejemplo, a la vez que miro ahora por la ventana los árboles y el lago, se me pide que describa lo que veo, la respuesta tendría una amplitud indefinida. No veo sólo esos árboles como árboles, sino como pinos, como semejantes a los pinos de California, pero diferentes en algunos aspectos, como semejantes en estos aspectos pero desemejantes en aquéllos, etc.

9. *El centro y la periferia*

Dentro del campo de la conciencia, necesitamos distinguir entre aquellas cosas que están en el centro de nuestra atención y las que están en la periferia. Somos conscientes de un vasto número de cosas a las que no estamos prestando atención o sobre las que no nos estamos concentrando. Por ejemplo: hasta este momento he estado concentrando mi atención en el problema filosófico de la descripción de la con-

ciencia, y no he estado prestando atención alguna a lo que siento respecto de la silla que está en contacto con mi espalda, a lo que me aprietan los zapatos que llevo puestos, o al ligero dolor de cabeza que me produce el haber bebido demasiado vino la otra noche. Sin embargo, todos estos fenómenos son parte de aquello de lo que, conscientemente, me doy por enterado. A menudo hablamos, en términos coloquiales, de tales rasgos de nuestra vida consciente como si fueran inconscientes, pero es un error decir, por ejemplo, que siento inconscientemente el roce de mi camisa contra mi piel en el sentido de que no soy consciente del crecimiento de las uñas de mis pies. Dicho brevemente: necesitamos distinguir entre la distinción consciente/inconsciente, por un lado, y la distinción centro de atención/periferia, por otro.

Consideremos otro ejemplo. Cuando iba conduciendo en mi coche esta mañana, la mayor parte de mi atención se dirigía hacia pensamientos filosóficos. Sin embargo, no es verdadero decir que conducía inconscientemente. El conducir inconscientemente me habría llevado a un desastre automovilístico. Yo estaba consciente durante todo el viaje, pero el centro de mi preocupación no era el tráfico ni la carretera, más bien lo eran los pensamientos sobre problemas filosóficos. Este ejemplo ilustra que es esencial distinguir entre diferentes niveles de atención dentro de los estados conscientes. Cuando conducía mi coche hacia la universidad esta mañana, mi nivel más elevado de atención se dirigía hacia los problemas filosóficos que me preocupaban. En un nivel inferior de atención, pero todavía a un nivel que puede describirse como *atención*, estaba prestando atención al conducir. Y de hecho, en algunas ocasiones, sucederían cosas que exigirían mi *atención completa*, tales como que tendría que dejar de pensar sobre filosofía y concentrar toda mi atención sobre la carretera. Además de esos dos niveles de atención, había también muchas cosas de las que, periféricamente, me daba por enterado, pero que no estaban próximas al centro de mi atención. Esto incluiría cosas tales como los árboles y las casas que estaban a la orilla de la carretera mientras pasaba, el roce del asiento del coche sobre mi espalda y del volante sobre mis manos, y la música que sonaba en la radio del coche.

Es importante darse cuenta de las distinciones correctas porque existe a menudo la tentación de decir que muchas cosas que están en la periferia de nuestra conciencia son en realidad inconscientes. Y esto es un error. Dreyfus (1991) cita frecuentemente el ejemplo de Heidegger del martillear del carpintero ducho en su oficio. El carpintero, cuando

martillea los clavos, puede estar pensando en su novia, o sobre su almuerzo, y puede no estar prestando atención al martillear. Pero es totalmente erróneo sugerir que está dando martillazos de manera inconsciente. A menos que sea un zombi total o una máquina inconsciente, es completamente consciente de su martillear, aunque esto no esté en el centro de su atención.

William James formuló una ley de la que es útil que nos acordemos. La expresa del modo siguiente: «La conciencia desaparece cuando no se la necesita». Pienso que se expresa mejor de la manera siguiente: «La atención desaparece cuando no se la necesita». Cuando, por ejemplo, me pongo por vez primera los zapatos, la presión y el roce de los zapatos están en el centro de mi conciencia; o cuando me siento en una silla, el roce de la silla está en el centro de mi conciencia. Pero el concentrarme en estas cosas no es realmente necesario para capacitarme para habérmelas con el mundo, y después de un rato, los rasgos de los zapatos y de la silla se retiran hacia la periferia de mi conciencia; ya no son el centro por más tiempo. Si tengo un clavo en mi zapato o si me caigo de la silla, entonces esas experiencias pasan a ser el centro de mi conciencia. Creo que aquello a lo que James se refiere es al centro y a la periferia de la conciencia, más bien que a la conciencia como tal.

10. *Condiciones límite*

Al reflexionar sobre el presente, no he tenido en ningún momento ningún pensamiento que tenga que ver con dónde estoy colocado, qué día del mes es hoy, en qué estación del año estoy, cuánto tiempo ha pasado desde que desayuné, cuáles son mi nombre y mi historia pasada, de qué país soy ciudadano, y así sucesivamente. Con todo, me parece que todo ello es parte de la situación, parte de la localización espaciotemporal y sociobiológica de mis estados conscientes presentes. Cualquier estado de conciencia está característicamente localizado de esta manera. Pero la localización puede no ser en sí el objeto de la conciencia, ni incluso en la periferia.

Una manera de darse cuenta de hasta qué punto los límites de la conciencia lo invaden todo es el de tener presente los casos en los que fallan. Hay, por ejemplo, un sentido de la desorientación que le sobreviene a uno cuando se da cuenta de repente de que es incapaz de acordarse del mes en que está, o de dónde está o de la hora que es.

11. *Estados de ánimo*

He mencionado antes que a menudo tenemos estados de ánimo que no son intencionales, aunque sean conscientes. Puedo estar en un estado de ánimo relajado o deprimido, un estado de ánimo alegre o abatido, y estos estados de ánimo no necesitan estar conscientemente dirigidos a ninguna condición de satisfacción intencional. Un estado de ánimo jamás constituye por sí mismo el contenido total de un estado consciente. Más bien, el estado de ánimo proporciona el tono o el color que caracteriza la totalidad de un estado consciente o de una secuencia de estados conscientes.

¿Estamos siempre en uno u otro estado de ánimo? La respuesta depende de la amplitud con que queramos interpretar la noción de estado de ánimo. Ciertamente, no estamos siempre en un estado de ánimo que tenga un nombre en un lenguaje como el castellano. En este momento, no estoy especialmente alegre ni especialmente deprimido; no estoy extasiado ni desesperado; tampoco estoy simplemente hablando por hablar. Con todo, me parece que hay en las experiencias que tengo en este momento lo que podría llamarse «tono». Y me parece que esto se puede asimilar apropiadamente a la noción general de estado de ánimo. El hecho de que mis experiencias presentes tengan un tono de alguna manera neutral no significa que no tengan en absoluto ningún tono. Es característico de los estados de ánimo el que invadan todas nuestras experiencias conscientes. Para la persona que está alegre, la visión del árbol, el paisaje y el cielo es una fuente de gran regocijo; para la persona que está desesperada, la misma visión produce sólo más depresión. Me parece que es característico de la vida consciente humana normal que estemos siempre en uno u otro estado de ánimo, y que este estado de ánimo invada todas nuestras formas conscientes de intencionalidad, aunque éste no sea, ni necesite ser, intencional.

Nada mejor que un cambio radical para que uno se dé cuenta de hasta qué punto el estado de ánimo lo invade todo. Cuando el estado de ánimo normal de una persona cambia radicalmente hacia arriba o hacia abajo, hacia una alegría o hacia una depresión inesperadas, uno se da cuenta de repente del hecho de que está siempre en un estado de ánimo u otro y que el estado de ánimo en el que uno está invade todos sus estados conscientes. Para mucha gente la depresión, desgraciadamente, es mucho más común que la alegría.

Mi conjetura es que tendremos una buena explicación neurobioló-

gica del estado de ánimo de manera más fácil que, digamos, las emociones. Los estados de ánimo lo invaden todo, son más bien simples, especialmente porque no tienen intencionalidad esencial y parece que debería haber una explicación bioquímica de algunos estados de ánimo. Tenemos ya fármacos que se usan para aliviar la depresión clínica.

12. *La dimensión placer/displacer*

Recuérdese que estamos considerando la totalidad de un estado consciente, una rebanada del flujo de la conciencia que sea lo suficientemente grande para tener la unidad y la coherencia que estoy intentando describir. Me parece que hay siempre, en tal porción, una dimensión de placer y displacer. Uno puede siempre plantear alguna de las preguntas de un inventario que incluye: «¿Era divertido o no?», «¿Te gustó o no?», «¿Cuando tenías dolor estabas desesperado, molesto, divertido, aburrido, extasiado, con naúseas, con asco, entusiasta, aterrado, irritado, encantado, feliz, infeliz, etc.?». Además hay muchas subdimensiones en la dimensión placer/displacer. Es posible, aunque excéntrico, aburrirse durante el éxtasis sexual y estar exultante mientras se padece dolor físico. Como sucede con el estado de ánimo, tenemos que evitar el error de suponer que las posiciones intermedias, y que, por lo tanto, carecen de nombre, no están en absoluto en la escala.

II. TRES ERRORES TRADICIONALES

Paso ahora a considerar tres tesis sobre los estados conscientes que, aunque se aceptan de manera bastante amplia, me parece que son, de acuerdo con una interpretación natural, falsas. Son las siguientes:

1. Todos los estados conscientes son autoconscientes.
2. La conciencia se conoce por medio de una facultad especial de introspección.
3. El conocimiento de nuestros estados conscientes es incorregible. No podemos equivocarnos sobre estos asuntos.

Consideremos estas tesis una por una.

1. *Autoconciencia*

Se argumenta[4] algunas veces que todo estado de conciencia es también un estado de autoconciencia; que es característico de los estados mentales conscientes que sean, por así decirlo, conscientes de sí mismos. No estoy seguro de qué hacer con esta afirmación, pero estoy seguro de que si la examinamos encontraremos que es o trivialmente verdadera o simplemente falsa.

Para empezar, necesitamos distinguir la noción ordinaria y no problemática de autoconciencia de la noción técnica del filósofo. En el sentido ordinario, hay claramente estados de conciencia en los que quizás soy consciente de mi propia persona, pero no soy necesariamente consciente de mis propios estados conscientes. Ilustraremos esto con algunos ejemplos.

Supongamos, en primer lugar, que estoy sentado en un restaurante comiendo un entrecot. En el sentido ordinario, no sería característicamente *auto*consciente en absoluto. Podría ser consciente de que el entrecot sabe bien, de que el vino con que lo acompaño es demasiado joven, de que las patatas están demasiado hechas, etc. Pero no hay autoconciencia.

Supongamos, en segundo lugar, que, de repente, me doy cuenta de que todo el mundo que está en el restaurante me está mirando fijamente. Podría preguntarme por qué están mirando embobados de esa manera hasta que descubro que mi distracción habitual me ha hecho la mala jugada de hacer que me olvide de ponerme los pantalones. Estoy allí, sentado en el restaurante, en calzoncillos. Tal circunstancia podría producir sentimientos que describiríamos normalmente como «autoconciencia aguda». Me doy cuenta de mi propia persona y del efecto que estoy produciendo en los demás. Pero incluso aquí mi autoconciencia no está dirigida a todos mis estados conscientes.

En tercer lugar, imaginemos que ahora estoy en el restaurante completamente vestido, y que concentro de repente mi atención en las experiencias conscientes que estoy teniendo en el restaurante al comer mi almuerzo y beber el vino. De repente me parece que, por ejemplo, me he estado revolcando inexcusablemente en una especie de autocomplacencia hiperestética y que he puesto mucho tiempo, esfuerzo y dinero en conseguir *esas* experiencias gastronómicas. De repente todo parece *excesivo.*

4. Por ejemplo, por David Woodruff Smith (1986).

Este me parece también un caso de autoconciencia en el sentido ordinario, pero difiere del segundo en que la autoconciencia se dirige a los estados de conciencia del agente mismo y no a su persona pública.

Ahora bien, en el caso ordinario de la autoconciencia, como se ejemplifica en los casos segundo y tercero, me parece simplemente falso que todo caso de conciencia sea un caso de autoconciencia. En el caso ordinario, la autoconciencia es una forma extraordinariamente sofisticada de sensibilidad y probablemente es poseída sólo por los seres humanos y algunas otras, muy pocas, especies.

Así pues, la afirmación de que toda conciencia involucra autoconciencia tiene que tener un sentido técnico. ¿Cuál es ese sentido? Hemos visto en nuestra exposición sobre la distinción entre el centro y la periferia que podemos siempre cambiar nuestra atención de los objetos que están en el centro de la conciencia hacia aquellos que están en la periferia, de modo que lo que era previamente periférico se convierta en central. De forma similar, parece que podemos siempre cambiar nuestra atención del *objeto* de la experiencia consciente a la *experiencia* misma. Podemos, por ejemplo, hacer siempre la jugada que hicieron los pintores impresionistas. Los pintores impresionistas produjeron una revolución en pintura cambiando su atención del objeto hacia la experiencia visual efectiva que tenían cuando miraban al objeto. Este es un caso de autoconciencia sobre el carácter de las experiencias. Me parece que podríamos obtener un sentido de «autoconciencia» donde es trivialmente verdadero que cualquier estado consciente es autoconsciente: en cualquier estado consciente podemos cambiar nuestra atención hacia el estado mismo. Por ejemplo, puedo concentrar mi atención no en la escena que está frente a mí, sino en la experiencia de mi ver esa misma escena. Y puesto que la posibilidad de ese cambio de atención estaba presente en el estado mismo podemos decir, en este sentido técnico muy especial, que todo estado consciente es autoconsciente.

Pero dudo mucho de que este sea el sentido que tienen presente los que afirman que toda conciencia es autoconciencia. Excepto en este sentido muy especial, me parece simplemente falso hacer esta afirmación.

2. *Introspección*

¿Se conocen los estados mentales mediante una capacidad especial, la capacidad de la introspección? En los capítulos anteriores he intenta-

do arrojar dudas sobre este punto de vista que es el prevalente tanto en filosofía como en el sentido común. Como en el caso de la autoconciencia, hay tanto una noción técnica como una de sentido común de la introspección. A menudo introspeccionamos, en sentido ordinario, nuestros propios estados conscientes. Supongamos, por ejemplo, que Sally desea saber si se casará o no con Jimmy, el cual acaba de proponérselo. Uno de sus procedimientos podría ser razonablemente examinar sus sentimientos de manera detallada. Y a esto, en castellano ordinario, se lo podría llamar una forma de introspección. Se hace a sí misma preguntas tales como: «¿Lo quiero realmente?», y si es el caso, «¿Cuánto?» «¿Cuáles son mis sentimientos más profundos respecto de él?», etc. El problema, según creo, no tiene que ver con el uso ordinario de la noción de introspección, sino con nuestro impulso como filósofos en tomar la metáfora literalmente. La metáfora sugiere que tenemos una capacidad de examinar nuestros propios estados conscientes, una capacidad modelada sobre la visión. Pero ese modelo o analogía es, seguramente, erróneo. En el caso de la visión tenemos una distinción clara entre el objeto visto y la experiencia visual que el perceptor tiene cuando percibe el objeto. Pero no podemos hacer esta distinción para el caso de los propios estados mentales conscientes. Cuando Sally vuelve su atención hacia dentro para introspeccionar sus sentimientos más profundos sobre Jimmy, no puede echarse hacia atrás para tener una vista mejor y dirigir su mirada al objeto que existe independientemente de sus sentimientos hacia Jimmy. Dicho brevemente: si por «introspección» queremos decir simplemente pensar sobre nuestros propios estados mentales, entonces no hay objeción alguna a la introspección. Esto sucede siempre y es crucial para cualquier forma de autoconocimiento. Pero si por «introspección» queremos decir que tenemos una capacidad especial, semejante a la visión sólo que con menos colores, que tenemos que *mirar dentro*, entonces me parece que no hay tal capacidad. No la puede haber porque el modelo de mirar dentro exige una distinción entre el objeto mirado y el mirarlo, y no podemos hacer esta distinción en el caso de los estados conscientes. Podemos dirigir un estado mental hacia otro estado; podemos pensar sobre nuestros pensamientos y sentimientos; y podemos tener sentimientos sobre nuestros pensamientos y sentimientos; pero nada de eso involucra ninguna facultad especial de introspección.

3. *Incorregibilidad*

Se dice a menudo que no podemos equivocarnos sobre los contenidos de nuestras propias mentes. De acuerdo con la concepción cartesiana tradicional de la mente, los informes en primera persona de los estados mentales son, de alguna manera, *incorregibles.* De acuerdo con este punto de vista, tenemos un cierto género de *autoridad de primera persona* en los informes sobre nuestros estados mentales. Se ha mantenido incluso que esta incorregibilidad es un signo seguro de que algo es mental (Rorty, 1970). Pero si se piensa un momento sobre ello, la afirmación de incorregibilidad parece obviamente falsa. Consideremos a Sally y a Jimmy. Sally podría llegar a darse cuenta más adelante que estaba simplemente equivocada cuando pensaba que estaba enamorada de Jimmy; que el sentimiento se había adscrito incorrectamente; se trataba sólo, de hecho, de una forma de apasionamiento. Y alguien que la conociese bien podría saber desde el principio que estaba equivocada.

Dados tales hechos, ¿por qué podría pensar alguien que era imposible que uno estuviese equivocado sobre los contenidos de sus propios estados mentales? ¿Por qué, para empezar, habría de suponer alguien que eran «incorregibles»? La respuesta tiene que ver quizás con la confusión entre la ontología subjetiva de lo mental y la certeza epistémica. Efectivamente, los estados mentales tienen una ontología subjetiva, como repetidamente he dicho a lo largo de este libro. Pero del hecho de que la ontología sea subjetiva no se sigue que uno no pueda equivocarse respecto de sus propios estados mentales. Todo lo que se infiere es que los modelos de error estándar, modelos basados en la distinción apariencia/realidad, no funcionan respecto de la existencia y la caracterización de los estados mentales. Pero estas no son las únicas formas posibles de estar equivocado respecto de un fenómeno. Todos sabemos a partir de nuestras propias experiencias que sucede a menudo que otra persona está en mejor posición de la que estamos nosotros para determinar si, por ejemplo, estamos realmente celosos, enfadados o con sentimientos de generosidad. Es verdad que el modo en el que me relaciono con mis estados mentales y, por lo tanto, el modo en que me relaciono con mis informes de mis estados mentales, es diferente del modo en que otras personas se relacionan con mis estados mentales. Y esto afecta al estatus de sus informes sobre mis estados mentales. Sin embargo, sus informes pueden ser más exactos que los míos.

¿En qué sentido exactamente se supone que tengo autoridad de pri-

mera persona sobre los contenidos de mi propia mente y por qué? Wittgenstein intentó valientemente, en las *Investigaciones filosóficas* (1953), eliminar *en absoluto* la idea de que debemos pensar en mis emisiones de primera persona como *informes o descripciones*. Si pudiésemos, como Wittgenstein sugería, pensar en ellas más bien como expresiones (*Aeusserungen*), entonces no serían en absoluto informes o descripciones y, por lo tanto, no habría cuestión alguna de autoridad. Cuando simplemente grito de dolor, no se plantea ninguna cuestión de autoridad, puesto que mi conducta de dolor era simplemente una reacción natural causada por el dolor, y no clase alguna de afirmación. Si mi decir «Tengo dolor» pudiese tratarse como una especie de grito ritualizado, una forma convencionalizada de conducta de dolor, entonces no se plantearía cuestión alguna sobre mi autoridad. Pienso que es justo decir que la solución que Wittgenstein intentó dar a este problema ha fallado. Hay efectivamente algunos casos en los que la conducta verbal respecto de los propios estados mentales se contempla más naturalmente como una forma de expresión del fenómeno mental más bien que como una descripción suya (por ejemplo, ¡Ay!), pero tenemos aún muchos casos en los que se está intentando dar un enunciado o descripción cuidadosos de los propios estados mentales y no se está dando simplemente expresión a ese estado. Ahora bien, ¿qué clase de «autoridad» tiene uno en esas emisiones y por qué?

Creo que el modo de captar lo que hay de especial en los informes de primera persona es preguntar por qué no pensamos que tenemos la misma autoridad especial sobre los objetos y estados de cosas del mundo que son *distintos* de nuestros estados mentales. La razón es que en nuestros informes sobre el mundo en general existe una distinción entre cómo las cosas nos parecen y cómo son realmente. Puede parecerme que hay una persona que se esconde en los arbustos que hay afuera, frente a mi ventana, cuando de hecho la apariencia era causada simplemente por el peculiar patrón de luz y sombra de los arbustos. Pero no hay distinción que pueda hacerse entre apariencia y realidad para cómo las cosas me parecen a mí. Realmente me parece que hay un hombre que está escondido entre los arbustos. El origen, dicho brevemente, de nuestra convicción de que hay una autoridad especial de primera persona reside simplemente en el hecho de que no podemos hacer la distinción convencional apariencia/realidad para las apariencias mismas. Esto plantea dos cuestiones. En primer lugar, ¿cómo es posible que podamos estar equivocados sobre nuestros propios estados mentales? ¿Cuál

es, por así decirlo, la *forma* del error que cometemos, si no es lo mismo que los errores apariencia/realidad que cometemos sobre el mundo en general? Y en segundo lugar, dado que las apariencias son parte de la realidad, ¿por qué no habríamos de ser capaces de hacer la distinción apariencia/realidad en el caso de las apariencias? Podemos comenzar a responder a la primera pregunta si exploramos alguno de los modos en que uno puede estar equivocado sobre si él mismo, por ejemplo, está enfadado o no. Dejando de lado la cuestión de los errores puramente lingüísticos, —esto es, dejando de lado los casos en los que una persona piensa que «enfadado» significa contento—, algunos casos típicos en los que uno da malas descripciones de sus propios fenómenos mentales son el autoengaño, la mala interpretación y la falta de atención. Los consideraré uno por uno.

Me parece bastante fácil «demostrar» la imposibilidad de autoengaño, pero el autoengaño es, a pesar de todo, un fenómeno bastante extendido y, por lo tanto, debe de haber algo erróneo en la demostración. La demostración procede del modo siguiente: para que *x* engañe a *y* debe de tener una creencia de que *p* y tiene que intentar con éxito inducir en *y* la creencia de que no *p*. Pero en el caso de que *x* es idéntico a *y*, parece como si *x* tuviese que producir en sí mismo la creencia autocontradictoria de que *p* y de que no *p*. Y esto parece imposible.

Sabemos, sin embargo, que el autoengaño es posible. Sin duda hay muchas formas de autoengaño, pero en una forma muy común el agente tiene un motivo o una razón para no admitir ante sí mismo que está en cierto estado mental. Puede avergonzarse del hecho de que está enfadado o de que odia a cierta persona o a cierta clase de gente. En tales casos, el agente se resiste simplemente a pensar conscientemente sobre algunos de sus estados psicológicos. Cuando el pensamiento de esos estados surge, inmediatamente piensa en el estado inverso al que desea mantener. Supóngase que odia a los miembros de un grupo minoritario, pero se avergüenza de este prejuicio y desea conscientemente no haber tenido este odio. Cuando se enfrenta con la evidencia de este prejuicio, simplemente rehúsa admitirlo, y de hecho, lo niega sincera y vehementemente. El agente tiene odio justamente con el deseo de no tener ese odio, esto es: con una forma de vergüenza sobre ese odio. Para reconciliar estas dos cosas, el agente evita conscientemente pensar sobre su odio y así es capaz de negarse sinceramente a admitir la existencia de ese odio cuando se enfrenta a la evidencia. Esta es seguramente una forma común de autoengaño.

Una segunda forma de «error» que puede cometerse respecto de los propios fenómenos mentales es la mala interpretación. Por ejemplo, en el punto álgido de una pasión una persona puede pensar que está enamorada; de hecho, piensa sinceramente que está enamorada, pero más tarde llega a darse cuenta de que estaba interpretando mal simplemente sus sentimientos. Para esta clase de caso es crucial la operación de la Red y el Trasfondo. Igual que una persona puede interpretar mal un texto si no logra ver cómo sus elementos se relacionan entre sí, y si no logra entender la operación de las circunstancias del Trasfondo en que el texto se compuso, del mismo modo una persona puede interpretar mal sus propios estados intencionales si no logra localizarlos correctamente de manera relativa al Trasfondo de capacidades mentales no representacionales. En tales casos no tenemos el modelo epistémico tradicional de hacer *inferencias* correctas sobre la base de *evidencia* insuficiente. No se trata de una cuestión de pasar de apariencia a realidad, sino más bien de localizar una pieza de un *puzzle* de manera relativa a todo un conjunto de piezas.

Finalmente, un caso bastante obvio de «error» respecto de los propios estados mentales es simplemente la falta de atención. En nuestras múltiples y caóticas ocupaciones de la vida no prestamos a menudo atención detallada a nuestros estados conscientes. Por ejemplo, una famosa política anunció recientemente en la prensa que se había equivocado al pensar que simpatizaba con los demócratas. Sin darse cuenta sus simpatías se habían desviado hacia los republicanos. Lo que tenemos en este caso es toda una Red de intencionalidad —cosas tales como las actitudes hacia la legislación, la simpatía hacia ciertas clases de políticos y la hostilidad hacia otras, reacciones a ciertos sucesos en política exterior, etc.— y esta Red había cambiado sin que ella se diese cuenta. En tales casos nuestros errores tienen que ver con la concentración de la atención, más bien que con la distinción tradicional entre apariencia y realidad.

III. Conclusión

Pienso que al menos dos errores, y quizás los tres, tienen un origen común en el cartesianismo. Los filósofos de la tradición cartesiana en epistemología querían que la conciencia proporcionase un fundamento para todo el conocimiento. Pero para que la conciencia nos dé una cier-

ta fundamentación del conocimiento, tenemos que tener primero cierto conocimiento de los estados conscientes; de ahí la doctrina de la incorregibilidad. Para conocer la conciencia con certeza, tenemos que conocerla por medio de alguna facultad especial que nos dé acceso directo a ella; de ahí la doctrina de la introspección. Y —aunque tengo menos confianza en esto como diagnosis histórica— si el yo ha de ser la fuente de todo conocimiento y significado, y esto ha de basarse en su propia conciencia, entonces es natural pensar que hay una conexión necesaria entre conciencia y autoconciencia; de ahí la doctrina de la autoconciencia.

En cualquier caso, diversos ataques recientes a la conciencia, tal como el de Dennett (1991), se basan en la suposición errónea de que si podemos mostrar que hay algo que es erróneo en la doctrina de la incorregibilidad o de la introspección, hemos mostrado que hay algo erróneo en la conciencia. La incorregibilidad y la introspección no tienen nada que ver con los rasgos esenciales de la conciencia. Son simplemente elementos de teorías filosóficas sobre la conciencia que están equivocados.

7. EL INCONSCIENTE Y SU RELACIÓN CON LA CONCIENCIA

El propósito de este capítulo es explicar las relaciones entre los estados mentales inconscientes y la conciencia. El poder explicativo de la noción del inconsciente es tan grande que no podemos habérnoslas sin él, pero la noción está lejos de ser clara. Esta falta de claridad tiene, como veremos, algunas consecuencias desafortunadas: diré también algo sobre la concepción freudiana de la relación entre la conciencia y el inconsciente, porque creo que, en la base, es incoherente. Haré un uso intenso de las distinciones entre epistemología, causación y ontología que expliqué en el capítulo 1.

I. EL INCONSCIENTE

Las generaciones precedentes —antes del siglo XX, dicho de manera aproximada— consideraban no problemática la noción de conciencia y un tanto intrigante y quizás autocontradictoria la noción de una mente inconsciente. Hemos dado la vuelta a los papeles. Después de Freud invocamos de manera rutinaria los fenómenos mentales inconscientes para explicar los seres humanos, y encontramos la noción de conciencia un tanto intrigante e incluso poco científica. Este cambio en el énfasis explicativo ha tomado formas diversas, pero la tendencia general en ciencia cognitiva ha sido la de introducir una cuña entre los procesos mentales subjetivos, conscientes, que no se consideran como un tema genuino de investigación científica, y aquellos procesos que se consideran como el tema genuino de la ciencia cognitiva y que, por lo tanto, tienen que ser objetivos. El tema general es que los procesos mentales inconscientes son más importantes que los conscientes. El enun-

ciado más radical de esto está contenido quizás en la afirmación de Lashley: «*Ninguna actividad de la mente es jamás consciente*» (las cursivas son de Lashley).[1] Otra versión extrema de este enfoque se encuentra en la afirmación de Ray Jackendoff (1987) de que hay, de hecho, dos «nociones de la mente», la «mente computacional» y la «mente fenomenológica».

Creo que a pesar de nuestra complacencia en el uso del concepto del inconsciente, no tenemos una noción clara de los estados mentales inconscientes, y mi primera tarea de clarificación va a ser la de explicar las relaciones entre el inconsciente y la conciencia. La afirmación que haré puede enunciarse con una sola oración: *La noción de estado mental inconsciente implica accesibilidad a la conciencia.* No tenemos noción alguna del inconsciente excepto como aquello que es potencialmente consciente.

Nuestra noción ingenua y preteórica de estado mental *inconsciente* es la idea de un estado mental consciente menos la conciencia. ¿Pero qué significa esto exactamente? ¿Cómo podríamos substraer la conciencia de un estado mental y quedarnos, con todo, con un estado *mental*? Desde Freud nos hemos ido acostumbrando a hablar de estados mentales inconscientes de tal manera que hemos perdido de vista el hecho de que la respuesta a esta pregunta no es, en absoluto, obvia. Con todo, resulta claro que no pensamos en el inconsciente bajo el modelo de lo consciente. Nuestra idea de un estado inconsciente es la idea de un estado mental que, simplemente, resulta ser aquí y ahora inconsciente, pero todavía lo entendemos bajo el modelo de un estado consciente, en el sentido de que pensamos en él como algo que es parecido a un estado consciente y como algo que, en algún sentido, podría haber sido consciente. Esto es claramente verdad, por ejemplo, en Freud, cuyas nociones de lo que él llama estados «preconscientes» e «inconscientes» se construyen sobre un modelo más bien simple de los estados conscientes (Freud, 1949, especialmente pp. 19-25). En su versión más ingenua el cuadro que se nos presenta es, más o menos, así: los estados mentales inconscientes de la mente son como peces que están en el fondo del mar. El pez que no podemos ver debajo de la superficie tiene

1. Lashley (1956). No creo que Lashley intente decir esto literalmente. Creo que quiere decir que los procesos mediante los que se producen los diversos rasgos de los estados conscientes no son jamás conscientes. Pero incluso esto es una exageración, y el hecho de que recurra a este tipo de hipérbole es revelador del tema que estoy intentando identificar.

exactamente la misma forma que cuando emerge. El pez no pierde sus formas al sumergirse bajo las aguas. Otro símil: los estados mentales inconscientes son como objetos almacenados en el oscuro desván de la mente. Esos objetos conservan siempre sus formas, incluso cuando no se pueden ver. Sentimos tentaciones de reírnos ante estos modelos tan simples, pero pienso que algo parecido a este cuadro es lo que subyace en nuestra concepción de los estados mentales conscientes, y es importante intentar ver lo que esta concepción tiene de correcto y de equivocado.

Como he mencionado antes, ha habido en las décadas recientes un esfuerzo bastante sistemático para separar la conciencia de la intencionalidad. La conexión entre las dos se ha ido perdiendo gradualmente, no sólo en la ciencia cognitiva, sino también en lingüística y en filosofía. Pienso que el motivo subyacente —y quizás inconsciente— de este impulso, que consiste en separar la intencionalidad de la conciencia, es que no sabemos cómo explicar la conciencia, y nos gustaría tener una teoría de la mente que no resultase desacreditada por el hecho de que carece de una teoría de la conciencia. La idea es tratar la intencionalidad «objetivamente», tratarla como si los rasgos subjetivos de la conciencia no importasen realmente. Por ejemplo, muchos funcionalistas conceden que el funcionalismo no puede «manejar» la conciencia (a esto se le denomina el problema de los *qualia*; véase el capítulo 2), pero piensan que este problema no tiene importancia alguna respecto de sus explicaciones de creencia, deseo, etc., puesto que estos estados intencionales no tienen ningún *quale*, ninguna cualidad consciente especial. Pueden tratarse como si fueran completamente independientes de la conciencia. De forma similar, tanto la idea de algunos lingüistas de que hay reglas de sintaxis que son psicológicamente reales pero totalmente inaccesibles a la conciencia, como la idea de algunos psicólogos de que hay inferencias complejas en la percepción que son procesos inferenciales psicológicos pero inaccesibles a la conciencia, implican una separación entre intencionalidad y conciencia. La idea en ambos casos no es que hay fenómenos mentales que sucede, por así decirlo, que son inconscientes, sino más bien que, de alguna manera, son *en principio* inaccesibles a la conciencia. No son el género de cosa que podría ser, o podría haber sido, consciente.

Pienso que estos desarrollos recientes son erróneos. Por profundas razones, nuestra noción de estado mental inconsciente es parásita de nuestra noción de estado consciente. Desde luego, en un momento dado

una persona puede estar inconsciente; puede estar dormida, en coma, etc., y, desde luego, muchos estados mentales no alcanzan nunca el nivel consciente. Y sin duda hay muchos que no pueden alcanzar el nivel consciente por una u otra razón —pueden ser demasiado dolorosos y, por lo tanto, pueden estar demasiado reprimidos para que, por ejemplo, pensemos en ellos. Sin embargo, no todo estado de un agente es un estado mental, y no todo estado del cerebro que funciona esencialmente en la *producción* de estados mentales es, en sí mismo, un fenómeno mental. Así pues, ¿qué es lo que hace que algo sea mental cuando no es consciente? Para que un estado sea un estado mental deben cumplirse ciertas condiciones. ¿Cuáles son?

Para explorar estas cuestiones, consideremos en primer lugar casos que son claramente mentales, aunque inconscientes, y contrastémoslos con casos que son «inconscientes» puesto que no son mentales en absoluto. Piénsese en la diferencia, por ejemplo, entre mi creencia (cuando no estoy pensando sobre ello) de que la torre Eiffel está en París, y la mielinización de los axones de mi sistema nervioso central. Hay un sentido en que ambas cosas son inconscientes. Pero hay una gran diferencia entre ellas: los estados estructurales de mis axones no podrían ser en sí mismos estados conscientes, puesto que no tienen nada de mental. Supongo por mor del argumento que la mielinización funciona esencialmente en la producción de mis estados mentales, pero incluso si los axones mielinizados fuesen ellos mismos objetos de experiencia, incluso si pudiera sentir internamente el estado de las cubiertas de mielina, con todo las estructuras efectivas no son estados mentales. No todo rasgo inconsciente de mi cerebro que (como la mielinización) funciona esencialmente en mi vida mental es un rasgo mental. Pero la creencia de que la torre Eiffel está en París es un estado mental genuino, incluso si sucede que es un estado mental que la mayor parte del tiempo no está presente en la conciencia. Así pues, hay en mí dos estados: mi creencia y la mielinización de mis axones; ambos tienen algo que ver con mi cerebro, y ninguno de los dos es consciente. Pero sólo uno es mental, y necesitamos clarificar qué lo hace mental y la conexión entre este rasgo —cualquiera que sea— y la conciencia. Precisamente para mantener clara la distinción, propongo llamar en este capítulo a fenómenos como la mielinización, que no están en absoluto en la línea de lo mental, fenómenos «no conscientes», y a fenómenos como los estados mentales en los que no estoy pensando o que se han reprimido «inconscientes».

Hay al menos dos constricciones en nuestra concepción de la intencionalidad de las que debe ser capaz de dar cuenta cualquier teoría del inconsciente. En primer lugar, tiene que ser capaz de dar cuenta de la distinción entre fenómenos que son genuinamente intencionales y aquellos que, en algunos aspectos, se comportan como si lo fuesen, pero de hecho no lo son. Esta es la distinción que establecí al final del capítulo 3 entre formas de intencionalidad *intrínseca* y *como-si*.[2] Y en segundo lugar, debe de ser capaz de dar cuenta del hecho de que los estados intencionales representan sus condiciones de satisfacción solamente bajo ciertos aspectos, y esos aspectos tienen que importar al agente. Mi creencia inconsciente de que la torre Eiffel está en París satisface ambas condiciones. El que yo tenga esa creencia es un asunto de intencionalidad intrínseca, y no un asunto de lo que cualquier otra persona elija decir de mí sobre cómo me comporto, o de qué género de postura se pueda adoptar respecto de mí. Y la creencia de que la torre Eiffel está en París representa sus condiciones de satisfacción bajo ciertos aspectos y no bajo otros. Es, por ejemplo, distinta de la creencia de que la estructura de acero más alta construida en Francia antes de 1900 se localiza en la capital de Francia, incluso suponiendo que la torre Eiffel es idéntica a la estructura de acero más alta construida en Francia antes de 1900, y París es idéntica a la capital de Francia. Podríamos decir que todo estado intencional tiene un cierto *contorno de aspecto*, y este contorno de aspecto es parte de su identidad, parte de lo que lo hace ser el estado que es.

II. EL ARGUMENTO A FAVOR DEL PRINCIPIO DE CONEXIÓN

Estos dos rasgos —el hecho de que un estado intencional inconsciente tenga que ser, a pesar de todo, intrínsecamente mental, y el hecho de que tenga que tener un cierto contorno de aspecto— tienen importantes consecuencias para nuestra concepción del inconsciente. Proporcionarán las bases para un argumento que muestra que sólo entendemos la noción de estado mental inconsciente como un contenido posible de conciencia, sólo como el género de cosa que, aunque no sea consciente, aunque sea quizás imposible traerlo al nivel de la conciencia por varias razones, es sin embargo el *género de cosa* que podría ser o haber

2. Véase también Searle (1980b, 1984b), y especialmente (1984a).

sido consciente. Llamaré a esta idea de que todos los estados intencionales inconscientes son en principio accesibles a la conciencia el principio de conexión, y voy a reproducir ahora el argumento en que se apoya con mayor detalle. En aras de la claridad, numeraré los pasos más importantes del argumento, aunque esto no quiere implicar que el argumento sea una simple deducción a partir de los axiomas.

1. *Hay una distinción entre intencionalidad* intrínseca *e intencionalidad* como-si; *sólo la intencionalidad intrínseca es genuinamente mental.* He argumentado con cierta extensión a favor de esta distinción que es más bien obvia, tanto en este libro como en los escritos que se han mencionado previamente, y no repetiré aquí los argumentos. Creo que la distinción es correcta y que el precio de abandonarla sería que todo sería mental, puesto que con relación a uno u otro propósito podría tratarse *como-si* lo fuese. Por ejemplo, el agua que cae colina abajo puede describirse *como-si* tuviese intencionalidad: *trata* de alcanzar la base de la colina *buscando* ingeniosamente la línea de menor resistencia, lleva a cabo *procesamiento de información, calcula* el tamaño de las rocas, el ángulo de inclinación, el empuje de la gravedad, etc. Pero si el agua es mental, entonces todo es mental.

2. *Los estados intencionales inconscientes son intrínsecos.* Cuando digo de alguien que está dormido que cree que George Bush es el presidente de los Estados Unidos, o cuando digo de alguien que está despierto que aborrece de manera inconsciente aunque reprimida a su padre, estoy hablando de manera completamente literal. No hay nada metafórico o del tipo *como-si* en estas atribuciones. Las atribuciones del inconsciente pierden su poder explicativo si no las tomamos literalmente.

3. *Los estados intencionales intrínsecos, ya sean conscientes o inconscientes, tienen siempre contornos de aspecto.* He venido utilizando el término técnico «contorno de aspecto», para señalar un rasgo universal de la intencionalidad. Puede explicarse como sigue: siempre que percibimos algo o pensamos sobre algo, lo hacemos siempre bajo unos aspectos y no bajo otros. Esos rasgos de aspecto son esenciales para el estado intencional; son parte de lo que lo hace el estado mental que es. El contorno de aspecto es más obvio si cabe en el caso de las percepciones conscientes: piénsese, por ejemplo, en ver un coche. Cuando se

ve un coche no se trata simplemente de que un objeto es registrado por el aparato perceptivo; más bien uno tiene una experiencia consciente del objeto desde un cierto punto de vista y con ciertos rasgos. Uno ve un coche como algo que tiene cierto contorno, que tiene cierto color, etc. Y lo que es verdad de las percepciones conscientes es verdad, de manera general, de los estados intencionales. Una persona puede creer, por ejemplo, que una estrella que está en el firmamento es la estrella de la mañana sin creer que es la estrella de la tarde. Una persona puede querer, por ejemplo, beber un vaso de agua sin querer beber un vaso de H_2O. Hay un número indefinidamente extenso de descripciones verdaderas de la estrella de la tarde y de vasos de agua, pero sólo se cree o se desea algo sobre estas cosas bajo ciertos aspectos y no bajo otros. Toda creencia y todo deseo, incluso todo fenómeno intencional, tiene un contorno de aspecto.

Obsérvese, además, que el contorno de aspecto tiene que interesar al agente. Es desde el punto de vista del agente desde el que él puede querer agua sin querer H_2O. En el caso de pensamientos conscientes, el modo en qué importa el contorno de aspecto viene dado porque constituye el modo en que el agente piensa sobre o experimenta los objetos sobre los que piensa o experimenta: puedo pensar, estando sediento, sobre las ganas que tengo de un trago de agua sin pensar en absoluto sobre su composición química. Puedo pensar en él *como* agua sin pensar en él *como* H_2O.

Está razonablemente claro cómo funciona esto para los pensamientos y las experiencias conscientes, ¿pero cómo funciona para los estados mentales inconscientes? Un modo de abordar nuestra cuestión es preguntar qué hecho sobre un estado mental inconsciente hace que tenga el particular contorno de aspecto que tiene, esto es: ¿qué hecho sobre él hace que sea el estado mental que es?

4. *El rasgo del aspecto no puede caracterizarse sólo, de manera exhaustiva o completa, en términos de predicados de tercera persona, conductistas, o incluso neurofisiológicos.* La evidencia conductista respecto de la existencia de estados mentales, incluyendo la pura evidencia respecto de la causación de la conducta de una persona, por completa que ésta sea, deja siempre indeterminado el carácter de aspecto de los estados intencionales. Habrá siempre un vacío inferencial entre los fundamentos *epistémicos* conductistas de la presencia del aspecto y la *ontología* del aspecto mismo.

Una persona puede, ciertamente, exhibir una conducta de búsqueda de agua, pero cualquier conducta de búsqueda de agua será también una conducta de búsqueda de H_2O. No hay manera, pues, de que la conducta, interpretada sin referencia a un componente mental, pueda constituir querer agua más bien que querer H_2O. Obsérvese que no es suficiente sugerir que podríamos conseguir que la persona en cuestión respondiese afirmativamente a la pregunta «¿Quieres agua?» y negativamente a la pregunta «¿Quieres H_2O?», puesto que las respuestas afirmativa y negativa son insuficientes para fijar el contorno de aspecto bajo el que tal persona interpreta la pregunta y la respuesta. No hay modo de determinar desde la conducta solamente si la persona en cuestión quiere decir mediante «H_2O» lo que yo quiero decir mediante «H_2O», y si quiere decir mediante «agua» lo que yo quiero decir mediante «agua». No hay cantidad alguna de hechos conductistas que constituyan el hecho de que alguien represente lo que quiere bajo un aspecto y no bajo otro. Este no es un asunto epistémico.

Es igualmente verdadero, aunque quizás de modo menos obvio, que ninguna acumulación de hechos neurofisiológicos bajo descripciones neurofisiológicas constituye hechos de aspecto. Incluso si tuviésemos una ciencia perfecta del cerebro, incluso si tal ciencia perfecta del cerebro nos permitiese colocar nuestro cerebroscopio en el cráneo de la persona en cuestión y ver que quería agua pero no H_2O, habría con todo que hacer todavía una inferencia; tendríamos que tener todavía alguna conexión legaliforme que nos capacitase para inferir a partir de nuestras observaciones de la arquitectura neural y de las descargas neuronales que eran realizaciones del deseo de agua y no del deseo de H_2O.

Puesto que los hechos neurofisiológicos son siempre causalmente suficientes para cualquier conjunto de estados mentales,[3] alguien con perfecto conocimiento causal podría ser capaz de hacer la inferencia de lo neurofisiológico a lo intencional al menos en aquel puñado de casos donde hay una conexión legaliforme entre los hechos especificados en términos neurales y los hechos especificados en términos intencionales. Pero incluso en esos casos, si es que hay alguno, hay *aún* una infe-

3. Para estos propósitos estoy contraponiendo «neurofisiológico» y «mental» pero, desde luego, de acuerdo con el punto de vista sobre las relaciones entre mente y cuerpo que he estado exponiendo a lo largo de este libro, lo mental es neurofisiológico a un nivel más elevado. Contrapongo lo mental y lo neurofisiológico como podrían contraponerse los seres humanos y los animales sin implicar por ello que la primera clase no está incluida en la segunda. No hay dualismo alguno implícito en mi uso de esta contraposición.

rencia, y la especificación de lo neurofisiológico en términos neurofisiológicos no es todavía una especificación de lo intencional.

5. *Pero la ontología de los estados mentales inconscientes, en el momento en que son inconscientes, consiste enteramente en la existencia de fenómenos puramente neurofisiológicos.* Imaginemos que alguien está profundamente dormido y no está soñando. Ahora bien, mientras que está en tal estado es verdadero decir de él que tiene cierto número de estados mentales inconscientes. Por ejemplo: cree que Denver es la capital de Colorado, Washington es la capital de los Estados Unidos, etc. ¿Pero *qué hecho sobre él hace que sea el caso que tiene estas creencias inconscientes*? Bien, los únicos hechos que podrían existir mientras está completamente inconsciente son hechos neurofisiológicos. Las únicas cosas que suceden en su cerebro inconsciente son secuencias de eventos neurofisiológicos que ocurren en las arquitecturas neuronales. En el momento en que los estados son totalmente inconscientes, simplemente no hay nada excepto estados y procesos neurofisiológicos.

Pero ahora parece que tenemos una contradicción. La ontología de la intencionalidad inconsciente consiste enteramente en fenómenos neurofisiológicos, objetivos, de tercera persona, pero a la vez los estados tienen un contorno de aspecto que no puede estar constituido por tales hechos, puesto que no hay contorno de aspecto alguno en el nivel de las neuronas y las sinapsis.

Creo que hay una única solución a este problema. La contradicción aparente se resuelve señalando que:

6. *La noción de estado intencional inconsciente es la noción de un estado que es un posible pensamiento o experiencia conscientes.* Hay una gran cantidad de fenómenos mentales inconscientes, pero hasta el punto en que son genuinamente *intencionales* tienen que preservar, en algún sentido, su contorno de aspecto incluso cuando son inconscientes, y el único sentido que podemos dar a la noción de que preservan su contorno de aspecto cuando son inconscientes es que son posibles contenidos de conciencia.

Esta es nuestra primera conclusión principal. Pero esta respuesta a nuestra primera pregunta da lugar a otra: ¿qué se quiere decir mediante «posible» en las dos oraciones previas? Después de todo, podría ser completamente imposible que el estado ocurriese conscientemente a

causa de una lesión cerebral, una represión u otras causas. Así pues, ¿en qué sentido exactamente tiene que ser un posible contenido de un pensamiento o de una experiencia? Esta cuestión conduce a nuestra siguiente conclusión, que es realmente una explicación adicional del paso 6, y que está implicada por 5 y 6 juntas:

7. *La ontología del inconsciente consta de rasgos objetivos del cerebro capaces de causar pensamientos conscientes subjetivos.* Cuando describimos algo como un estado intencional inconsciente, estamos caracterizando una *ontología* objetiva en virtud de su capacidad *causal* de producir conciencia. Pero la existencia de estos rasgos causales es consistente con el hecho de que, en cualquier caso, dados sus poderes causales, pueden bloquearse por alguna otra causa que interfiera como, por ejemplo, la represión psicológica o el daño cerebral.

La posibilidad de interferencia por parte de diversas formas de patología no altera el hecho de que cualquier estado intencional inconsciente es la clase de cosa que es, en principio, accesible a la conciencia. Puede ser inconsciente no sólo en el sentido de que *sucede* que no es consciente aquí y ahora, sino también en el sentido de que por una razón u otra el agente *no podría* simplemente traerlo a la conciencia, pero tiene que ser la *clase de cosa* que puede traerse a la conciencia puesto que su ontología es la de una neurofisiología caracterizada en términos de su capacidad para causar conciencia.

Paradójicamente, el mentalismo ingenuo de mi punto de vista sobre la mente lleva a un género de análisis disposicional de los fenómenos mentales inconscientes; sólo que no se trata de una disposición a «comportarse», sino una «disposición» —si esta es realmente la palabra correcta— a tener pensamientos conscientes, incluyendo en esto pensamientos conscientes manifestados en la conducta. Esto es paradójico, incluso irónico, puesto que la noción de explicación disposicional de lo mental se introdujo precisamente para desembarazarse de la apelación a la conciencia; y estoy en efecto intentando poner patas arriba esta tradición argumentando que las creencias inconscientes son ciertamente estados disposicionales del cerebro, pero disposiciones a producir pensamientos y conducta conscientes. Esta clase de adscripción disposicional de capacidades causales nos es muy familiar a partir del sentido común. Cuando, por ejemplo, decimos de una substancia que es lejía o que es veneno, estamos adscribiendo a una ontología química una capacidad causal disposicional de pro-

ducir ciertos efectos. De forma similar, cuando decimos de alguien que está inconsciente que cree que Bush es presidente, estamos adscribiendo a una ontología neurobiológica la capacidad causal disposicional de producir ciertos efectos, a saber: pensamientos conscientes con contornos de aspecto específicos. El concepto de intencionalidad inconsciente es entonces el de *latencia* relativa a su *manifestación* en la conciencia.

Para resumir: el argumento a favor del principio de conexión era en cierto modo complejo, pero su línea subyacente era completamente simple. Pregúntate sólo a ti mismo que hecho del mundo se corresponde con tus afirmaciones. Cuando se hace una afirmación sobre intencionalidad inconsciente, no hay hechos que tengan que ver con el caso excepto los neurofisiológicos. No hay aquí nada más que estados y procesos neurofisiológicos describibles en términos neurofisiológicos. Pero los estados intencionales, conscientes o inconscientes, tienen contornos de aspecto, y no hay contorno de aspecto alguno en el nivel de las neuronas. Así pues, el único hecho sobre las estructuras neurofisiológicas que corresponde a la adscripción de contorno de aspecto intrínseco es el hecho de que el sistema tiene la capacidad de producir estados y procesos conscientes allí donde se manifiestan esos contornos de aspecto específicos.

El cuadro general que emerge es este. Lo único que ocurre en mi cerebro son procesos neurofisiológicos, algunos conscientes, algunos inconscientes. De entre los procesos neurofisiológicos inconscientes unos son mentales, otros no. La diferencia entre ellos no reside en la conciencia, puesto que, por hipótesis, ninguno de ellos es consciente; la diferencia reside en que los procesos mentales son candidatos para la conciencia, puesto que son capaces de causar estados conscientes. Y esto es todo. Toda mi vida mental está alojada en el cerebro. ¿Pero qué es en mi cerebro mi «vida mental»? Estas dos cosas solamente: estados conscientes y aquellos estados y procesos que —dadas las circunstancias correctas— son capaces de generar estados conscientes. Llamemos a estos estados que, en principio, son accesibles a la conciencia «someramente inconscientes» y a aquellos que son inaccesibles, incluso en principio, «profundamente inconscientes». La conclusión principal de este capítulo es que no hay estados mentales profundamente inconscientes.

II. Dos objeciones al principio de conexión

Quiero discutir dos objeciones. La primera de ellas la he pensado 'o mismo, aunque muchas otras personas[4] me han dado también difeentes versiones de ella; la segunda se debe a Ned Block.

Primera objeción: supongamos que tuviésemos una ciencia perfeca del cerebro. Supongamos, por ejemplo, que pudiésemos poner nuesro cerebroscopio en el cráneo de alguien y ver que quería agua. Suongamos ahora que la configuración «Quiero-agua» del cerebro fuese niversal. Las personas quieren agua si y sólo si tienen esa configuraión. Esto es, desde luego, una fantasía completa de ciencia ficción, ero démosla por buena. Supongamos ahora que encontramos un subector de la población que tuviese exactamente esa configuración pero ue no pudiese «en principio» provocar en la conciencia deseo alguno le agua. En ellos no hay nada patológico; así es precisamente como sus erebros han sido construidos. Ahora bien, si esto es posible —¿y por ué no?— hemos encontrado un contraejemplo al principio de coneión, puesto que hemos encontrado un ejemplo de un deseo inconsiente de agua que, en principio, resulta imposible de provocar en la onciencia.

Me gusta el ejemplo, pero no pienso que sea de provocar. En las iencias definimos característicamente fenómenos superficiales en términos de sus microcausas; podemos definir los colores, por ejemplo, n términos de longitud de onda de cierto número de milmillonésimas le metro. Si tuviésemos una ciencia perfecta del cerebro de la clase que emos imaginado, podríamos ciertamente identificar estados mentales on sus microcausas en la neurofisiología del cerebro. Pero —y esto es l punto crucial— la redefinición funciona como una identificación de un fenómeno mental inconsciente sólo hasta el punto en que continuamos suponiendo que la neurofisiología del inconsciente está todavía, por así decirlo, rastreando el fenómeno mental consciente correcto con el contorno de aspecto correcto. Así pues, la dificultad reside en el uso de la expresión «en principio». En el caso que hemos imaginado, la neurofisiología del «Quiero-agua» es capaz, de hecho, de causar la experiencia consciente. Obtuvimos el ejemplo que va en primer lugar sólo bajo este supuesto. Los casos que hemos imaginado son simplemente casos en los que se ha producido algún bloqueo de alguna clase.

4. Específicamente, David Armstrong, Alison Gopnik y Pat Hayes.

Son semejantes a los ejemplos de «vista ciega» de Weiskrantz, sólo que sin la patología. Pero en los fenómenos en cuestión no hay nada «en principio» inaccesible a la conciencia, y esta es la razón por la que no es un contraejemplo al principio de conexión.

Segunda objeción. El argumento tiene la consecuencia de que no podría haber un zombi intencional totalmente inconsciente. ¿Por qué no podría haberlo? Si algo semejante es posible —¿y por qué no?— entonces el principio de conexión entraña una proposición falsa y es, por lo tanto, falso.

Efectivamente, no podría haber un zombi intencional, y el famoso argumento de Quine a favor de la indeterminación de la traducción (Quine, 1960, cap. 2) nos proporciona, sin proponérselo, la prueba. Para un zombi, a diferencia de un agente consciente, no hay simplemente hecho objetivo alguno respecto de qué contornos de aspecto tienen exactamente sus pretendidos estados intencionales. Supóngase que construimos un zombi «buscador de agua». Ahora bien, ¿qué hecho sobre el zombi hace que él, o ella, esté buscando esto bajo el aspecto «agua» y no bajo el aspecto «H_2O»? Obsérvese que no sería suficiente para responder a esta pregunta señalar que podríamos programar al zombi para decir: «Quiero, ciertamente, agua, pero no quiero H_2O», puesto que esto sólo fuerza la discusión un paso atrás: ¿qué hecho sobre el zombi hace que sea el caso que mediante «agua» quiere decir lo que nosotros queremos decir mediante «agua», y mediante «H_2O», quiere decir lo que nosotros queremos decir mediante «H_2O»? Incluso si complicásemos su conducta para tratar de responder a esta pregunta, siempre habrá modos alternativos de interpretar su conducta verbal que serán consistentes con todos los hechos sobre conducta verbal, pero que dan atribuciones inconsistentes de significado e intencionalidad al zombi. Y, como Quine ha mostrado detalladamente de manera muy elaborada, el problema no es que no podemos saber con seguridad que el zombi quiso decir, por ejemplo, «conejo» como opuesto a «estado en la historia de la vida de un conejo», o «agua» como opuesto a «H_2O», sino que no hay hecho objetivo alguno sobre lo que el zombi quiso decir. Pero donde no hay hecho objetivo alguno sobre el contorno de aspecto, no hay contorno de aspecto, no hay intencionalidad. Quine, podríamos decir, tiene una teoría del significado apropiada para zombis que se manifiestan verbalmente. Pero nosotros no somos zombis y nuestras emisiones tienen, en algunas ocasiones al menos, significados determinados con contornos de aspecto determinados, lo mismo

que nuestros estados intencionales tienen a menudo contenidos intencionales determinados con contornos de aspecto determinados (Searle, 1987). Pero todo esto presupone conciencia.

IV. ¿PODRÍA HABER DOLORES INCONSCIENTES?

Quiero ilustrar adicionalmente el principio de conexión imaginando un caso en el que tendríamos un uso de la noción de «dolor inconsciente». Normalmente, no pensamos en dolores inconscientes y creo que mucha gente aceptaría la noción cartesiana de que para que algo sea un dolor genuino tiene que ser consciente. Pero pienso que es fácil evocar intuiciones contrarias. Considérese lo siguiente: sucede comúnmente entre personas que sufren dolores crónicos, digamos dolores crónicos de espalda, que algunas veces el dolor les hace difícil irse a la cama. Y de hecho, una vez que se duermen, hay ocasiones en las que, durante la noche, *su condición provocar que se despierten.* Ahora bien, ¿cómo describiremos exactamente esos casos? Supongamos, por mor de este ejemplo, que los pacientes están completamente inconscientes durante el sueño; no tienen conciencia de dolor alguno. ¿Diremos entonces que durante el sueño no tienen realmente dolor alguno, sino que el dolor comenzó cuando se despertaron y que fueron despertados por procesos neurofisiológicos que normalmente causarían dolor, pero que no causaron dolor puesto que entonces los pacientes estaban dormidos? ¿O diremos, por otra parte, que el dolor, esto es, el dolor mismo, continuó tanto antes como durante como después de su sueño, pero que no eran conscientes del dolor mientras estaban dormidos? Mis intuiciones encuentran lo segundo completamente natural, de hecho probablemente más natural que lo primero. Sin embargo, lo importante es ver que aquí no hay un problema substantivo involucrado. Estamos simplemente adoptando un vocabulario alternativo para describir el mismo conjunto de hechos. Pero consideremos ahora el segundo vocabulario: de acuerdo con este vocabulario, decimos que el dolor fue consciente durante algún tiempo, que a continuación fue inconsciente, que después fue consciente de nuevo. Mismo dolor, diferentes estados de conciencia. Podríamos incrementar nuestro impulso a hablar de esta manera si encontrásemos que la persona, aunque completamente inconsciente, hace movimientos corporales durante el sueño que servían para proteger la parte dolorosa de su cuerpo.

Ahora bien, ¿cuál es exactamente la ontología del dolor cuando es completamente inconsciente? La respuesta me parece completamente obvia. Lo que nos inclina a decir que el dolor continuó existiendo incluso inconscientemente es que había un proceso neurofisiológico subyacente que era capaz de generar un estado consciente y capaz de generar conducta apropiada a alguien que tenía ese estado consciente. Y en el ejemplo, tal como se ha descrito, es esto lo que ha pasado.

Pero si ahora estoy en lo correcto, entonces es difícil ver cómo puede haber substancia fáctica alguna en las viejas discusiones entre los freudianos y sus adversarios respecto de si los estados mentales inconscientes existen realmente. Si se acepta mi argumento hasta aquí, entonces no soy capaz de ver cómo podría ser algo más que un asunto puramente terminológico, diferente sólo en complejidad del problema sobre la existencia de dolores inconscientes tal como lo he descrito aquí. Un bando insistía en que hay realmente estados *mentales inconscientes*; el otro insistía en que si eran realmente *mentales* tienen que ser *conscientes*. ¿Pero qué hechos del mundo se supone que corresponden a estas dos diferentes afirmaciones?

La evidencia que los freudianos aducían involucraba historias causales, conducta y admisiones conscientes por parte del agente —todo lo cual parecía interpretable solamente bajo la suposición de un estado mental inconsciente, que era exactamente igual a un estado consciente, excepto por el hecho de que era inconsciente. Considérese un caso típico. A una persona que está hipnotizada se le da una indicación posthipnóptica de que tiene que andar a gatas por el suelo después de salir del trance hipnótico. Más adelante, cuando está consciente, da una justificación completamente extraña, pero aparentemente racional, de su conducta. Dice, por ejemplo: «Creo que puedo haber perdido mi reloj por aquí, por el suelo», y acto seguido se pone a andar a gatas. Ahora bien, suponemos, creo que con buenas razones, que está obedeciendo inconscientemente la orden, que inconscientemente intenta andar a gatas por el suelo puesto que el hipnotizador le dijo que lo hiciese, y que la razón que da para su conducta no es, en absoluto, la razón real.

Pero suponiendo que es totalmente inconsciente de sus motivos reales ¿cuál se supone que es, aquí y ahora, la ontología del inconsciente? Para repetir nuestra primera pregunta, ¿qué *hecho* corresponde a la atribución del estado mental inconsciente en el momento en que el agente está actuando por una razón de la que es totalmente inconsciente? Si realmente el estado es totalmente inconsciente, entonces los únicos he-

chos son la existencia de estados neurofisiológicos capaces de dar lugar a pensamientos conscientes y a la clase de conducta apropiada para el que tiene estos pensamientos.

Algunas veces puede haber diversos pasos inferenciales entre el estado mental inconsciente latente y la intencionalidad consciente manifiesta. Así, se nos dice, el adolescente que se rebela en contra de la autoridad de la escuela está motivado inconscientemente por el odio hacia su padre. Pero de nuevo, como en el caso de la hipnosis, tenemos que preguntar ¿cuál se supone que es la ontología del inconsciente cuando es inconsciente? Y en este caso, como en el caso de la hipnosis, la atribución de un contorno de aspecto específico a lo inconsciente tiene que implicar que hay en la neurofisiología una capacidad de producir un pensamiento consciente con este mismo contorno de aspecto.

Una vez que se ve que la descripción de un estado mental como «inconsciente» es la descripción de una ontología neurofisiológica en términos de su capacidad causal de producir pensamientos y conducta conscientes, entonces parece que no puede haber substancia fáctica alguna en la pregunta ontológica ¿existen realmente los estados mentales inconscientes? Todo lo que esta pregunta puede significar es: ¿son capaces los estados neurofisiológicos del cerebro *no conscientes* de dar lugar a pensamientos conscientes y a las clases de conducta apropiadas para alguien que tenga esos pensamientos? Naturalmente, ninguna de las partes piensa en el problema de este modo, pero quizás parte de la intensidad del debate se derivaba del hecho de que lo que parecía un problema ontológico directo —¿existen los estados mentales inconscientes?— no era realmente un debate ontológico en absoluto.

Si tengo razón en esto, los viejos argumentos freudianos —que incluyen toda la evidencia resultante del hipnotismo, las neurosis, etc.— no son tanto conclusivos o inconclusivos como fácticamente vacíos. El problema no es menos importante por ser conceptual o terminológico, pero deberíamos entender que no es un problema fáctico sobre la existencia de entidades mentales que no son ni psicológicas ni conscientes.

V. La posición de Freud sobre el inconsciente

Quiero concluir este capítulo comparando mi concepción del inconsciente y su relación con la conciencia con la de Freud. De acuerdo con mi punto de vista, dentro de nuestros cráneos hay una masa de neu-

ronas empotradas en células gliales, y algunas veces este vasto e intrincado sistema es consciente. La conciencia está causada por la conducta de elementos de nivel inferior, presumiblemente en los niveles neuronales, sinápticos y columnares, y como tal es un rasgo de nivel superior del sistema total. Lo que digo no implica que conciencia y neurofisiología sean algo simple. Ambos asuntos me parecen inmensamente complejos y la conciencia, en particular, aparece, como hemos visto, en una gran variedad de modalidades: percepción, emoción, sentimiento, dolores, etc. Pero de acuerdo con mi punto de vista, esto es todo lo que pasa dentro del cerebro: procesos neurofisiológicos y conciencia. Hablar de conciencia, de acuerdo con mi explicación, es simplemente hablar de capacidades causales de la neurofisiología para causar estados y conducta consciente.

Hasta aquí mi posición. ¿Qué sucede con la de Freud? Donde yo veo adscripciones verdaderas de vida mental inconsciente que corresponden a una ontología neurofisiológica objetiva, si bien descrita en términos de su capacidad para causar fenómenos mentales subjetivos conscientes, Freud[5] ve esas adscripciones como algo que se corresponde con estados mentales que existen como estados mentales aquí y ahora. Esto es: Freud piensa que nuestros estados mentales inconscientes existen como inconscientes y, a la vez, como estados intencionales intrínsecos ocurrentes incluso cuando son inconscientes. ¿Puede hacer que este cuadro sea coherente? He aquí lo que dice: todos los estados mentales son «inconscientes en sí mismos». Y traerlos a la conciencia es simplemente algo parecido a percibir un objeto (1915, reimpreso en 1959, vol. 4, especialmente pp. 140 y ss.). Así pues, la distinción entre estados mentales conscientes e inconscientes no es una distinción entre dos géneros de estados mentales, o incluso una distinción entre dos diferentes modos de existencia de los estados mentales, sino que más bien todos los estados mentales son inconscientes en sí mismos (*an sich*) y lo que llamamos «conciencia» es sólo un modo de percepción de estados que son inconscientes en su modo de existencia. Es como si los estados mentales inconscientes fuesen realmente algo así como muebles que están en el desván de la mente y, para traerlos a la conciencia, subiésemos al desván y los iluminásemos con el destello de nuestra percepción. Igual que los muebles «en sí mismos» no se

5. Ignoro, en esta discusión, la distinción de Freud entre preconsciente e inconsciente. Para los presentes propósitos llamo a ambos «inconsciente».

ven, del mismo modo los estados mentales son «en sí mismos» inconscientes.

Es posible que esté interpretando mal a Freud, pero no puedo encontrar, o inventar, una interpretación coherente de su teoría. Incluso si dejamos aparte los estados conscientes de la percepción y nos limitamos a los estados intencionales proposicionales como las creencias y los deseos, me parece que la teoría es incoherente en, al menos, dos aspectos. En primer lugar, no puedo hacer coherente su explicación de la ontología con lo que sabemos sobre el cerebro y, en segundo lugar, no puedo formular una versión coherente de la analogía entre percepción y conciencia.

Esta es la primera dificultad: supongamos que tengo una serie de estados mentales inconscientes. Cuando estoy completamente inconsciente, lo único que sucede en mi cerebro son procesos neurofisiológicos que ocurren en arquitecturas neuronales específicas. Así pues, ¿qué hecho respecto de esos procesos y arquitecturas neurofisiológicas se supone que *constituye* el que sean estados mentales inconscientes? Obsérvense los rasgos que los estados mentales inconscientes tienen que tener *qua* estados mentales. En primer lugar, tienen que tener contorno de aspecto; y en segundo lugar tienen que ser, en algún sentido, «subjetivos», puesto que son *mis* estados mentales. Es fácil ver cómo los estados conscientes satisfacen estas condiciones —tales estados se experimentan como teniendo contorno de aspecto. Es más difícil, pero aún es posible, ver cómo las satisfacen los estados inconscientes si pensamos en la ontología del inconsciente del modo que he sugerido —como una neurofisiología ocurrente capaz de causar estados y eventos conscientes. ¿Pero cómo puede tener la neurofisiología no consciente contorno de aspecto y subjetividad aquí y ahora? La neurofisiología admite, de hecho, diferentes niveles de descripción, pero ninguno de esos niveles de descripción neurofisiológicos objetivos —que van de la microanatomía de la hendidura sináptica a órganos molares más extensos como el hipocampo— es un nivel de contorno de aspecto o de subjetividad.

Freud piensa aparentemente que, además de cualesquiera rasgos neurofisiológicos que mi cerebro pueda tener, hay también algún nivel de descripción en el que mis estados mentales inconscientes, aunque completamente inconscientes, tienen todos y cada uno de los rasgos de mis estados mentales conscientes, incluyendo la intencionalidad y la subjetividad. El inconsciente tiene todo lo que tiene el consciente, *sólo*

que sin la conciencia. Pero no ha hecho inteligible qué eventos podrían ocurrir en el cerebro, además de los eventos neurofisiológicos, para constituir la subjetividad y la intencionalidad inconscientes.

La *evidencia* que Freud nos da a favor de la existencia del inconsciente es invariablemente la de que el paciente emprende una conducta que es *como si* tuviese un cierto estado mental, pero puesto que sabemos independientemente que el paciente no tiene ningún estado mental consciente de este tipo, Freud postula un estado mental inconsciente como la causa de la conducta. Un verificacionista tendría que decir que el único significado que se puede postular aquí es que el paciente se comporta de tal y tal manera y que tal conducta estaría causada normalmente por un estado consciente. Pero Freud no es un verificacionista. Piensa que hay algo aquí que causa la conducta que no es neurofisiológico, pero que tampoco es consciente. No puedo hacer consistente esto con lo que sabemos sobre el cerebro, y es difícil interpretarlo excepto como una posición que implica dualismo, puesto que Freud postula una clase de fenómenos mentales no neurofisiológicos; esto parece constituir un abandono por su parte del primitivo proyecto de una psicología científica (1895).

¿Qué sucede con la analogía entre conciencia y percepción? Una vez que se adopta el punto de vista de que los estados mentales son *en sí mismos* mentales y *en sí mismos* inconscientes, entonces no va a ser fácil explicar cómo encaja la conciencia en el cuadro. Parece como si el punto de vista de que los estados mentales son en sí mismos inconscientes tiene la consecuencia de que la conciencia es totalmente extrínseca, de que no es una parte esencial de ningún estado o evento consciente. Me parece que Freud acepta esta consecuencia, y la analogía entre conciencia y percepción es un modo de intentar que la conciencia encaje dentro del cuadro, dada la consecuencia de que la conciencia es un rasgo extrínseco, no esencial de cualquier estado consciente. Una vez que se formula la teoría del inconsciente, la analogía con la percepción parece inevitable. Para dar cuenta del hecho de la conciencia junto con la teoría del inconsciente, nos vemos forzados a postular que la conciencia es un género de percepción de estados y eventos que en su naturaleza intrínseca son inconscientes.

Pero esta solución nos lleva de Guatemala a Guatepeor. Como hemos visto en nuestra exposición de la introspección, el modelo de la percepción funciona bajo el supuesto de que hay una distinción entre el objeto percibido y el acto de la percepción. Freud necesita esta suposi-

ción para dar cuenta de la consecuencia de que la conciencia es extrínseca, de que, por ejemplo, esta instancia de pensamiento consciente podría haber existido sin la conciencia. Intentemos tomar la analogía seriamente. Supongamos que veo una bicicleta. En tal situación perceptiva, tenemos una distinción entre el objeto percibido y el acto de percepción. Si dejo de lado la percepción me quedo con la bici; si dejo de lado la bici, me quedo con una percepción que no tiene objeto, por ejemplo, con una alucinación. Pero son precisamente estas distinciones las que no podemos hacer en el caso de un pensamiento consciente. Si intento separar el pensar consciente de esta instancia de pensamiento, pongamos por caso, de que Bush es presidente, no me queda nada. Si intento separar la ocurrencia de la instancia de pensamiento del pensarlo conscientemente, no logro separar nada. La distinción entre el acto de percibir y el objeto percibido no se aplica a los pensamientos conscientes.

Además, parece que caemos en un círculo vicioso si mantenemos que el fenómeno de provocar en la conciencia estados inconscientes consiste en percibir previamente fenómenos mentales inconscientes que, en sí mismos, son inconscientes. Pues surge entonces la pregunta: ¿qué pasa con el acto de percibir —es un fenómeno mental? Si es así, tiene que ser «en sí mismo» inconsciente, y parecería que para que uno fuese consciente de ese acto necesitaría algún acto de nivel superior de percibir mi acto de percibir. No estoy seguro de esto, pero tiene todo el aspecto de una amenaza de regreso al infinito.

Una dificultad final que tiene esta analogía con la percepción es la siguiente: la percepción funciona bajo el supuesto de que el objeto percibido ejerce un impacto causal sobre mi sistema nervioso, que causa la experiencia que tengo de él; así, cuando toco o siento algo, el objeto de la percepción causa una cierta experiencia. ¿Pero cómo podría funcionar posiblemente esto en el caso en que el objeto percibido fuese en sí mismo una experiencia inconsciente?

Para resumir: me parece que hay dos objeciones a la explicación freudiana. Una: no tenemos una noción clara de cómo se supone que encaja la ontología del inconsciente con la ontología de la neurofisiología. Dos: no tenemos una noción clara de cómo aplicar la analogía perceptiva a la relación entre consciente e inconsciente; parece, además, que caemos en el absurdo y en un regreso al infinito si intentamos tomarla en serio.

VI. Los restos del inconsciente

¿Qué queda del inconsciente? He dicho anteriormente que nuestra noción preteórica e ingenua del inconsciente era parecida a las nociones del pez en el mar o de muebles en el oscuro desván de la mente. Mantienen sus contornos aunque sean inconscientes. Pero ahora podemos ver que esas imágenes son inadecuadas en principio puesto que se basan en la idea de una realidad mental constante que aparece y, a continuación, desaparece. Pero la creencia sumergida, a diferencia del pez sumergido, no puede mantener su contorno consciente cuando es inconsciente; pues la única realidad ocurrente de ese contorno es el contorno de los pensamientos conscientes. La imagen ingenua de los estados inconscientes confunde la capacidad causal de causar un estado intencional consciente con el estado consciente mismo, esto es: confunde la latencia con su manifestación. Es como si pensásemos que la botella de veneno de la alacena tuviese que estar envenenando algo durante todo el tiempo para poder ser realmente veneno. Para repetirlo, *la ontología del inconsciente es estrictamente la ontología de una neurofisiología capaz de generar la conciencia.*

La conclusión final que quiero extraer de esta discusión es que no tenemos noción unificada alguna del inconsciente. Hay, al menos, cuatro nociones diferentes.

En primer lugar, hay atribuciones metafóricas —*como-si*— de intencionalidad al cerebro, que no han de tomarse literalmente. Por ejemplo, podríamos decir que la medula quiere mantenernos vivos; de este modo nos hace que sigamos respirando, incluso cuando dormimos.

En segundo lugar, existen los casos freudianos de deseos, creencias, etc., someramente inconscientes. Es mejor pensar en ellos como casos conscientes reprimidos, puesto que están siempre burbujeando en la superficie, aunque a menudo de una manera disfrazada. La noción freudiana del inconsciente es, en su comportamiento lógico, muy distinta de la noción de la ciencia cognitiva en el aspecto crucial de que los estados mentales freudianos inconscientes son potencialmente conscientes.

En tercer lugar, están los casos (relativamente) poco problemáticos de fenómenos mentales someramente inconscientes que da la casualidad de que no forman parte del contenido de mi conciencia en un determinado punto temporal. Así, la mayor parte de mis creencias, deseos, preocupaciones y recuerdos no están presentes en mi conciencia en un

momento dado, por ejemplo en el momento presente. Sin embargo, todos ellos son *potencialmente* conscientes en el sentido que he explicado (si lo entiendo correctamente, esto es lo que Freud quiso decir con «preconsciente» como opuesto al «inconsciente» [Freud, 1949]).

En cuarto lugar, se supone que hay una clase de fenómenos mentales intencionales profundamente inconscientes que no son sólo inconscientes sino que son, en principio, inaccesibles a la conciencia. No sólo no tenemos evidencia alguna a favor de su existencia, sino que la postulación de su existencia viola una constricción lógica de la noción de intencionalidad.

8. CONCIENCIA, INTENCIONALIDAD Y EL TRASFONDO

I. Introducción al Trasfondo

El propósito de este capítulo es explicar las relaciones entre conciencia e intencionalidad por un lado y, por otro, las capacidades, habilidades y saber-cómo general que hacen posible el funcionamiento de nuestros estados mentales. Llamo colectivamente a esas capacidades, etc., el «Trasfondo», con «T» mayúscula para dejar claro que uso la palabra como término técnico. Puesto que he desarrollado en algunos aspectos importantes mis puntos de vista sobre el Trasfondo desde que escribí *Intencionalidad* (1983), explicaré también los cambios y las motivaciones que he tenido para hacerlos.

Al principio de los setenta comenzé a investigar los fenómenos que más tarde llamé «el Trasfondo» y también a desarrollar una tesis que llamo «la hipótesis del Trasfondo». La tesis era originalmente una afirmación sobre el significado literal (Searle, 1978), pero creo que lo que se aplica al significado literal se aplica también al significado que intenta comunicar el hablante e, incluso, a todas las formas de intencionalidad, ya sean lingüísticas o no lingüísticas. La tesis del Trasfondo es, simplemente, esta: los fenómenos intencionales tales como significados, comprensiones, interpretaciones, creencias, deseos y experiencias funcionan sólo dentro de un conjunto de capacidades de Trasfondo que no son en sí intencionales. Otra manera de enunciar esta tesis es decir que toda representación, ya sea en el lenguaje, en el pensamiento o en la experiencia, sólo tiene éxito al representar dado un conjunto de capacidades no representacionales. En mi jerga técnica, los fenómenos intencionales sólo determinan *condiciones de satisfacción* con relación a un conjunto de capacidades que no son intencionales. Así

pues, el mismo estado intencional puede determinar diferentes condiciones de satisfacción, dadas diferentes capacidades de Trasfondo, y un estado intencional no determinará condiciones de satisfacción de ningún tipo a menos que se aplique con relación a un Trasfondo apropiado.

Para desarrollar adicionalmente esta tesis, necesito repetir una distinción que he hecho anteriormente entre el Trasfondo y la Red. En general, resulta imposible para los estados intencionales determinar aisladamente condiciones de satisfacción. Para tener una creencia o un deseo, tengo que tener toda una Red de otras creencias y deseos. Así, por ejemplo, si quiero hacer una buena comida en un restaurante de la ciudad, tengo que tener un gran número de otras creencias y deseos tales como las creencias de que hay un restaurante en la ciudad, que los restaurantes son la clase de establecimiento donde se sirven comidas, que las comidas son la clase de cosa que se puede comprar y comer dentro de los restaurantes a ciertas horas del día pagando una cierta cantidad de dinero, y así sucesivamente de manera más o menos indefinida. Sin embargo, el problema es este: incluso si tuviese la paciencia de hacer una lista de todas las demás creencias que constituyen la Red que da sentido a mi deseo de hacer una buena comida en un restaurante, todavía me queda el problema que me planteaba mi deseo inicial, a saber: que el contenido de la intencionalidad, por así decirlo, no se autointerpreta. Aún está sujeto a un rango indefinido de aplicaciones diferentes. Por lo que respecta al contenido intencional efectivo de mi deseo, es posible tener este mismo contenido y aplicarlo en un número indefinido de modos diferentes unos de otros e inconsistentes entre sí. ¿Qué constituye exactamente comer? ¿Qué constituye una comida? ¿Qué constituye un restaurante? Todas esas nociones están sujetas a interpretaciones diferentes, y esas interpretaciones no se fijan por el contenido del estado intencional mismo. Además de la Red, necesitamos postular un Trasfondo de capacidades que no son parte de la Red. O más bien, la totalidad de la Red necesita un Trasfondo, puesto que los elementos de la Red ni se autointerpretan ni se autoaplican.

La tesis del Trasfondo (en la que estoy incluyendo ahora las afirmaciones sobre la Red) constituye una tesis muy fuerte. Incluye al menos lo siguiente:

1. Los estados intencionales no funcionan autónomamente. No determinan aisladamente las condiciones de satisfacción.

2. Cada estado intencional requiere para su funcionamiento una Red de otros estados intencionales. Las condiciones de satisfacción se determinan sólo de manera relativa a la Red.

3. Incluso la Red no es suficiente. La Red sólo funciona de manera relativa a un conjunto de capacidades de Trasfondo.

4. Esas capacidades no son y no pueden ser tratadas como meros estados intencionales o como parte del contenido de algún estado intencional particular.

5. El mismo contenido intencional puede determinar diferentes condiciones de satisfacción (tales como las condiciones de verdad) y con relación a algún Trasfondo no determina ninguna en absoluto.

Para pensar en el Trasfondo ingenuamente, piénsese en la figura de Wittgenstein del hombre andando cuesta arriba. Podría interpretarse como un hombre deslizándose hacia atrás cuesta abajo. No hay nada que sea interno a la figura, incluso interpretada como una representación figurativa de un hombre en esa situación que nos fuerce a la interpretación que encontramos natural. La idea del Trasfondo es que lo que funciona para la figura funciona para la intencionalidad en general.

En el siglo pasado la clase de fenómeno que llamo «Trasfondo» fue reconocido por un número muy diferente de filósofos con compromisos muy distintos. Nietzsche no fue ciertamente el primero que reconoció el fenómeno, pero fue uno de los que más consciente fue de esta contingencia: el Trasfondo no tiene por qué ser del modo que es. No hay prueba alguna al efecto de que el Trasfondo que tenemos tenga que ser necesariamente el que es. La obra del último Wittgenstein es en gran parte sobre el Trasfondo.[1] Entre los escritores contemporáneos, me parece que la noción de Bourdieu (1990) de *habitus* está estrechamente relacionada con mi noción de Trasfondo.

En este capítulo bosquejaré en primer lugar un argumento a favor de la tesis del Trasfondo y, a continuación, intentaré justificar la postulación de los fenómenos del Trasfondo como una categoría separada para la investigación. En segundo lugar, volveré a enunciar la tesis del Trasfondo a la luz de la discusión de las relaciones entre conciencia, el inconsciente y la intencionalidad que se han presentado en el capítulo 7. En tercer lugar, expondré diversas implicaciones de la tesis del

1. Especialmente *Sobre la certeza* (1969), que creo que es uno de los mejores libros sobre el tema.

Trasfondo; en particular, intentaré evitar las diversas malas comprensiones y malas concepciones que me parece que ha generado el hecho de darse cuenta de la existencia del Trasfondo. En cuarto lugar, empezaré una explicación general del Trasfondo.

II. ALGUNOS ARGUMENTOS A FAVOR DE LA HIPÓTESIS DEL TRASFONDO

En obras anteriores (Searle, 1978, 1980c, 1983, 1990) he presentado argumentos a favor de estas cinco tesis, y no quiero repetirlos aquí. Sin embargo, para dar una visión general de las tesis que presento, bosquejaré alguna de las consideraciones que más me impresionan. El modo más simple de ver que la representación presupone un Trasfondo no representacional de capacidades es examinar la comprensión de oraciones. Lo bueno de empezar con oraciones viene dado por el hecho de que son objetos sintácticos bien definidos, y las lecciones que pueden aprenderse de ellas pueden aplicarse generalmente a los fenómenos intencionales. El punto número 5 nos da la cuña para entrar en el argumento: el mismo significado literal determinará condiciones de satisfacción diferentes, por ejemplo, diferentes valores de verdad, con relación a diferentes suposiciones de Trasfondo, y algunos significados literales no determinarán condiciones de verdad a causa de la ausencia de presuposiciones de Trasfondo apropiadas. Además (punto 4), esas suposiones de Trasfondo ni están ni pueden estar incluidas en el significado literal. Así, por ejemplo, si se consideran las ocurrencias de la palabra «corta» en «Sam corta la hierba», «Sally corta el pastel», «Bill corta la tela», «Él corta su piel», se verá que la palabra «corta» significa lo mismo en cada una de ellas. Esto se muestra, por ejemplo, por el hecho de que la reducción de la conjunción funciona para las ocurrencias de este verbo con esos objetos directos. Puede decirse, por ejemplo: «General Electric ha inventado un nuevo aparato que corta hierba, corta pasteles, corta tela y corta piel». Se pueden simplemente eliminar las últimas tres ocurrencias de «corta» y escribir «General Electric ha inventado un nuevo aparato que corta hierba, pasteles, tela y piel». Obsérvese que la palabra «corta» difiere en esas ocurrencias de sus ocurrencias metafóricas genuinas. Si digo «Sally corta con Bill», «La CNN corta sus emisiones mañana», «El rector corta la calefacción debido al plan de austeridad», en cada uno de los casos la palabra «corta» tiene un uso no literal. De nuevo, la reducción de la conjunción nos muestra esto. Si

digo: «General Electric ha inventado un aparato que corta hierba, pasteles, tela y piel» y a continuación añado «y con Bill, emisiones y suministros de calefacción», todo ello se convierte en un mal chiste. Así pues, las emisiones contienen la ocurrencia literal del verbo «corta», pero esta palabra, de acuerdo con una interpretación normal, se interpreta de maneras diferentes en cada oración. Se puede ver esto también si uno se imagina la correspondiente versión imperativa de esas emisiones. Si digo «Corta la hierba» y corres fuera y te pones a apuñarla con un cuchillo, o si digo «Corta el pastel» y te precipitas sobre él con una cortadora de césped, entonces hay aquí un sentido perfectamente ordinario en el que no has hecho lo que se te ha pedido que hicieses.

La lección que hay que aprender de estos ejemplos es la siguiente: la misma expresión literal puede hacer la misma contribución a la emisión literal de una gran variedad de oraciones y con todo, aunque esas oraciones se comprendan literalmente —no hay cuestión alguna de metáfora, ambigüedad, actos de habla indirectos, etc.—, la expresión será interpretada diferentemente en las diferentes oraciones. ¿Por qué? Porque cada oración se interpreta teniendo en cuenta un Trasfondo de capacidades humanas (habilidades para tomar parte en ciertas prácticas, saber-cómo, modos de hacer cosas, etc.), y esas capacidades fijarán diferentes interpretaciones, incluso si el significado literal de las expresiones permanece constante.

Ahora bien, ¿por qué es este un resultado importante? Bien, de acuerdo con nuestras explicaciones estándar del lenguaje, el significado de una oración es una función composicional de los significados de sus partes componentes y su disposición sintáctica en la oración. Así pues, entendemos la oración «Juan ama a María» de modo diferente al que entendemos la oración «María ama a Juan» precisamente a causa de la aplicación de la composicionalidad. Además, somos capaces de entender oraciones porque se componen de elementos significativos, elementos cuyos significados son asunto de convención lingüística. Así pues, el principio de composicionalidad y la noción de significado literal son absolutamente esenciales para una explicación coherente del lenguaje. Sin embargo, aunque necesarios para una explicación del lenguaje, sucede que no son suficientes. Además, necesitamos postular un Trasfondo no representacional.

Es tentador pensar que este argumento descansa sobre la ambigüedad, sobre casos marginales, etc. Pero esto es un error. Una vez que se

ha logrado que todo sea completamente explícito, una vez que se han eliminado todas las ambigüedades estructurales y léxicas, el problema del Trasfondo todavía se plantea. Esto puede verse si uno se da cuenta de que los esfuerzos progresivos de precisión no son suficientes para eliminar la necesidad del Trasfondo. Supóngase que entras en un restaurante y pides una comida. Supóngase que digo, hablando literalmente, «Tráigame un entrecot con patatas fritas». Aunque la emisión se diga y se entienda literalmente, el número de posibles malas interpretaciones es estrictamente ilimitado. Doy por sentado que no servirán la comida a mi domicilio o a mi lugar de trabajo. Doy por sentado que el entrecot no estará encerrado en hormigón ni petrificado. No se lo meterá en mis bolsillos ni se me tirará a la cabeza. Pero ninguna de estas suposiciones se hacía explícita en la emisión literal. La tentación es pensar que podría hacerlas completamente explícitas añadiéndoles restricciones adicionales haciendo mi encargo original más preciso. Pero esto es también un error. En primer lugar, es un error porque no hay límite al número de adiciones que tendría que hacer al encargo original para bloquear posibles malas interpretaciones y, en segundo lugar, cada una de las adiciones está sujeta ella misma a diferentes interpretaciones.

Otro argumento a favor del Trasfondo es este: hay oraciones perfectamente ordinarias del castellano y de otros lenguajes naturales que son ininterpretables. Entendemos todos los significados de las palabras, pero no entendemos la oración. Así, por ejemplo, si uno oye la oración «Sally corta la montaña», «Bill corta el sol», «Joe corta el lago», o «Sam corta el edificio», se encontrará perplejo respecto de lo que esas oraciones pueden significar. Si alguien te ha dado una orden así: «Vete y corta la montaña», realmente no sabes qué hacer. Sería fácil inventar una práctica de Trasfondo que fijase una interpretación literal de cada una de esas oraciones, pero sin tal práctica, no sabemos cómo aplicar el significado literal de la oración.

En la lingüística reciente hay algún reconocimiento de los problemas del Trasfondo (véanse, por ejemplo, los artículos de Robyn Carston y François Récanati que aparecen en Davis, 1991), pero los análisis que he visto sólo tocan la superficie del problema. Por ejemplo, una discusión bastante común se refiere a las relaciones entre el significado literal de la oración emitida, el contenido de lo que dice el hablante, y lo que el hablante implica al hacer la emisión. Así, por ejemplo, en la oración «He desayunado», el significado literal de la oración no hace referencia alguna al día de la emisión, pero nosotros interpretaríamos

normalmente esa emisión en el sentido de que su contenido es que el hablante ha desayunado *hoy*, esto es: el día de la emisión. Entonces, «He desayunado» contrasta con «He estado en el Tibet», una emisión que no nos comunica si he estado en el Tibet hoy. O considérese otra oración bastante discutida: «Sally le dio a John la llave y él abrió la puerta». Una emisión de esta oración comunicaría normalmente que *primero* Sally dio a John la llave, y *después* él abrió la puerta y la abrió con la llave. Hay muchas discusiones sobre los mecanismos mediante los cuales se comunica este contenido adicional, dado que no está codificado en el significado literal de la oración. La sugerencia, seguramente correcta, es que el significado oracional subdetermina, al menos hasta cierto punto, lo que el hablante dice cuando emite la oración. Ahora bien, la afirmación que estoy haciendo es esta: el significado oracional subdetermina *radicalmente* el contenido de lo que se dice. Considérense los ejemplos siguientes. Nadie interpretaría «He tenido la gripe» por analogía con «He tenido gemelos». Esto es, dado nuestro Trasfondo, nadie interpretaría que la oración significa «Acabo de dar a luz a la gripe», pero obsérvese que no hay absolutamente nada en el contenido semántico de la oración que bloquee esta interpretación, o incluso que nos obligue a la interpretación de que he *padecido* la gripe. Es muy fácil, aunque obsceno, imaginar una cultura en la que las dos interpretaciones de «He tenido...» estuviesen invertidas. Problemas similares surgen para cualquier oración. Considérese «Sally dio a John la llave y él abrió la puerta». No hay absolutamente nada en el contenido semántico literal de esta oración que bloquee la interpretación: «John abrió la puerta con la llave echando la puerta abajo; la llave medía ocho metros, estaba hecha de acero, y pesaba cien kilos». No hay nada que bloquee la interpretación: «John abrió la puerta con la llave tragándose la puerta y la llave e introduciendo la llave en la cerradura por medio de las contracciones peristálticas de su intestino». Desde luego, tales interpretaciones serían algo completamente disparatado, pero no hay nada en el contenido semántico de la oración, interpretada por sí misma, que bloquee estas descabelladas interpretaciones.

¿Hay algún modo en el que podamos dar cuenta de todas esas intuiciones sin una afirmación tan extrema como la tesis del Trasfondo? Bien, intentémoslo. Una idea, debida a François Récanati,[2] es la siguiente.

2. En una discusión.

Cualquier situación efectiva admite un número infinito de descripciones verdaderas, de modo que cualquier representación lingüística será incompleta. Si alguien «corta» el pastel pasando por encima de él un cortacésped, es verdadero decir «Él corta el pastel». Pero nos sorprendería bastante encontrarnos con esta oración que da cuenta de este evento. Nuestra sorpresa, sin embargo, no tiene nada que ver con la semántica, comprensión, etc. Tenemos simplemente un conjunto de expectativas basado en la inducción, y el informe, aunque verdadero, era incompleto puesto que dejaba fuera una explicación de cómo difiere el cortar del modo que nosotros esperaríamos normalmente.

Récanati me dice que no está de acuerdo con este punto de vista, pero yo lo encuentro importante y desafiante, de modo que lo quiero considerar con más profundidad. La sugerencia es: el significado literal fija las condiciones de verdad aisladamente, pero está acompañado por un sistema de expectativas, y este sistema funciona a la vez que el significado literal. El problema real sugerido por los ejemplos es que una vez que se eliminan de una oración todas las ambigüedades genuinas, nos quedamos todavía con vaguedad e incompletitud. Pero el hecho de que los significados sean *complementados* con un conjunto de expectativas habituales añade precisión y completitud adicionales a la comprensión. Así pues, no diríamos:

El significado literal sólo determina condiciones de verdad con relación a un Trasfondo.

Más bien diríamos:

El significado literal (dejando de lado la indexicalidad y otros rasgos dependientes del contexto) determina las condiciones de verdad absolutamente y de modo aislado. Pero los significados literales son vagos, y las descripciones literales son siempre incompletas. Se añade una mayor comprensión y completitud complementando el significado literal con suposiciones y expectativas colaterales. Así, por ejemplo, cortar es cortar lo hagas como lo hagas, pero esperamos que la hierba se corte de una manera y los pasteles de otra. Así, si alguien dice: «Ve y corta la montaña», la respuesta correcta no es «No lo entiendo». ¡Naturalmente que entiendes esa oración castellana! Más bien la respuesta correcta es «¿Cómo quieres que la corte?».

Pienso que este es un argumento poderoso y atractivo. Las respuestas que daría son dos. En primer lugar, si el problema fuera un problema de incompletitud, entonces deberíamos, en principio, acercarnos a la completitud añadiendo más oraciones adicionales. Pero no podemos. Como he señalado antes, cada oración que añadimos está sujeta a malas comprensiones adicionales a menos que esté fijada por el Trasfondo. En segundo lugar, si se supone una ruptura radical entre significado literal y «supuestos» colaterales, entonces uno debería de ser capaz de aplicar el significado literal sin importar cuáles sean los supuestos. Pero no se puede. Así, por ejemplo, la aplicación de la palabra «corta» sólo procede teniendo en cuenta la suposición de que algunos objetos del mundo son sólidos y admiten penetración por medio de la presión física de los instrumentos. Sin esta suposición no puedo interpretar la mayor parte de las ocurrencias de «corta». Pero esta suposición no es parte del significado literal. Si lo fuese, la introducción de dispositivos para cortar por medio de rayos láser involucraría un cambio en el significado de la palabra y, ciertamente, no lo involucra. Además, puedo imaginar usos literales de «corta» en un universo donde esta suposición es falsa. Podemos imaginarnos un conjunto de capacidades de Trasfondo en la que «Corta el lago» es algo perfectamente claro.

Creo que si se desarrollase completamente este argumento, se podría mostrar que si se postula una ruptura total entre significado literal y Trasfondo, uno se encontraría con un estilo de escepticismo kripkeano-wittgensteniano (Kripke, 1982), puesto que entonces uno sería capaz de decir cualquier cosa y querer decir mediante ella cualquier cosa.[3] Si se hace una ruptura radical entre significado y Trasfondo entonces, por lo que respecta al significado, cualquier cosa vale; pero esto implica que la comprensión normal ocurre sólo relativamente a un Trasfondo. No estoy, sin embargo, intentando demostrar ninguna tesis general sobre el escepticismo semántico.

Mis respuestas a esta objeción son, en primer lugar, que el problema no es la incompletitud, puesto que los esfuerzos por completar la descripción no sirven de ayuda. En algún sentido ni siquiera se empieza con esos esfuerzos, puesto que cada oración adicional sólo añade formas adicionales de incompletitud. Y en segundo lugar, si se postula

3. La respuesta correcta a este tipo de escepticismo consiste, creo, en explicar el papel del Trasfondo en el significado y la comprensión (Searle, inédito).

una situación totalmente desprovista de presuposiciones de Trasfondo, entonces no se puede fijar ninguna interpretación determinada.

Una segunda cuestión, planteada también por Récanati, es esta: ¿cuál es el argumento a favor de generalizar desde el significado literal todas las formas de intencionalidad? El único «argumento» que ofrecería es que es útil tener una taxonomía que capture nuestra intuición de que hay un ajuste entre pensamiento y significado. Por ejemplo, quiero capturar nuestra intuición ordinaria de que el hombre que tiene la creencia de que Sally corta el pastel tiene una creencia con exactamente el mismo contenido proposicional que la aserción literal «Sally corta el pastel». Puesto que estamos utilizando los términos técnicos «Trasfondo» e «intencionalidad», el uso ordinario no decidirá la disputa. Si se usa la noción de contenido intencional de tal manera que el significado literal sea una expresión de contenido intencional, entonces se sigue que las constricciones de Trasfondo se aplican por igual a ambos. Puedo imaginar otras taxonomías, pero ésta es la que me parece que funciona mejor.

Un buen modo de observar el Trasfondo es tener en cuenta los casos en que se produce algún fallo. Un ejemplo ilustrará lo que quiero decir. Un filósofo visitante vino a Berkeley y asistió a algunos seminarios sobre el Trasfondo. No estaba muy convencido por los argumentos. Un día tuvo lugar un pequeño terremoto. Esto le convenció porque, como me dijo más tarde, no había tenido, antes de ese momento, ninguna creencia, convicción o hipótesis de que la tierra no se mueve; simplemente la había dado por sentada. El punto es que «dar algo por sentado» no tiene por qué ser el nombre de un estado intencional completamente paralelo a creer o plantear una hipótesis.

Un paso crucial para entender el Trasfondo es ver que uno puede comprometerse con la verdad de una proposición sin tener estado intencional alguno con esa proposición como contenido.[4] Puedo, por ejemplo, estar comprometido con la proposición de que los objetos son sólidos sin que tenga de ningún modo, ni explícita ni implícitamente, creencia o convicción alguna a tal efecto. Pero bien, ¿cuál es entonces el sentido de compromiso que está involucrado aquí? Al menos este: no puedo, de una manera que sea consistente con mi conducta, negar esa proposición. No puedo, mientras estoy sentado en esta silla, mientras

4. Esto representa un cambio respecto del punto de vista que mantenía en Searle, 1991. William Hirstein me convenció de ello.

estoy apoyado sobre esta mesa y con los pies descansando sobre este suelo, negar de manera consistente que los objetos son sólidos, puesto que mi conducta presupone la solidez de esos objetos. Es en este sentido en el que mi conducta intencional, una manifestación de mis capacidades de Trasfondo, me compromete con la proposición de que los objetos son sólidos, incluso si no he necesitado formarme creencia alguna respecto de la solidez de los objetos.

Además, es importante ver que el Trasfondo no afecta meramente a problemas relativamente sofisticados tales como la interpretación de las oraciones, sino a rasgos fundamentales tales como aquellos que constituyen las bases formales de todo lenguaje. Por ejemplo, damos por sentado el hecho de que nuestro uso actual del lenguaje identifica instancias fonéticas y grafémicas del mismo tipo sintáctico, en virtud de contornos fonéticos y grafémicos, pero es importante ver que esto es una práctica contingente basada en capacidades de Trasfondo contingentes. En lugar de un lenguaje en el que la secuencia «Francia», «Francia», «Francia» involucra tres diferentes ocurrencias de la misma unidad sintáctica, podríamos imaginar fácilmente un lenguaje en el que el significado no va ligado a un tipo identificado fonética o grafemáticamente, sino a la secuencia numérica de ocurrencias-instancias del tipo. Así, por ejemplo, la primera vez que aparece en un discurso, la inscripción «Francia» podría usarse para hacer referencia a Francia, pero la segunda vez se refiere a España, la tercera a Italia, etc. La unidad sintáctica no es aquí la palabra en el sentido tradicional, sino una secuencia de inscripciones de instancias. Lo mismo sucede con los sistemas de oposición de los que los estructuralistas están tan orgullosos: el aparato de caliente como opuesto a frío, norte a sur, varón a mujer, vida a muerte, este a oeste, arriba a abajo, etc., están todos ellos basados en el Trasfondo. No hay nada inevitable en la aceptación de estas oposiciones. Sería fácil imaginarse unos seres para los que el este se opone naturalmente al sur, para los que sería ininteligible oponer este a oeste.

III. La Red es parte del Trasfondo

Intentaré ahora enunciar exactamente cómo mi presente punto de vista sobre las relaciones entre conciencia, inconsciente e intencionalidad, tal como se formularon en el capítulo anterior, produce una modificación —y, espero, una mejora— de mi concepción previa del Tras-

fondo. De acuerdo con mi punto de vista previo, yo pensaba en la mente como algo que contenía un inventario de estados mentales. En cualquier momento dado, algunos de ellos son conscientes y otros son inconscientes. Por ejemplo, podría pensar conscientemente que Bush es presidente, o podría tener inconscientemente esa creencia, una ocurrencia de una instancia de esa misma creencia, incluso cuando estoy totalmente dormido. Pero la conciencia no era esencial a los fenómenos mentales, ni incluso a las experiencias perceptivas, como los experimentos de Weiskrantz parecen mostrar.

De acuerdo con este punto de vista, algunos fenómenos que podrían enunciarse como creencias parece que se describen de modo poco natural si se los enuncia así. Efectivamente, tengo una creencia inconsciente de que Bush es presidente, cuando no estoy pensando sobre ello, pero parece que no tengo una creencia inconsciente en este sentido de que, por ejemplo, los objetos son sólidos. Simplemente, me comporto de tal manera que doy por sentada la solidez de los objetos. La solidez de los objetos no es parte de mis presuposiciones de Trasfondo; no es, en absoluto, un fenómeno intencional, a menos que se convierta en tal como parte, por ejemplo, de alguna investigación teórica.

Pero este modo de pensar las cosas me plantea algunas dificultades. ¿Cuál es la base de la distinción entre el Trasfondo y la Red? Bien, pidiendo la cuestión, puedo decir que el Trasfondo consta de fenómenos que no son estados intencionales, y la Red es una red de intencionalidad; pero ¿cómo se supone que debe delinearse exactamente esta distinción si se nos dice, por ejemplo, que mi creencia inconsciente de que Bush es presidente es parte de la Red y mi presuposición de que los objetos son sólidos es parte del Trasfondo? ¿Qué sucede con la creencia de que George Bush lleva ropa interior o de que tiene dos orejas? ¿Son parte también de mi Red consciente? Estamos cometiendo un error al plantear la pregunta de este modo. Y debería sernos obvio. De acuerdo con el punto de vista de la mente como algo que contiene un inventario de estados mentales, tiene que haber un error categorial al intentar trazar una línea entre Red y Trasfondo, puesto que el Trasfondo consta de un conjunto de capacidades, y la Red no es en absoluto un conjunto de capacidades, sino de estados intencionales.

Pienso ahora que el error real era suponer que existe un inventario de estados mentales, algunos conscientes, algunos inconscientes. Tanto el lenguaje como la cultura tienden a forzarnos a adoptar esta concepción. Pensamos en la memoria como un almacén de proposiciones e

imágenes, como un género de gran biblioteca o archivo de representaciones. Pero deberíamos pensar en la memoria más bien como un *mecanismo* para generar las realizaciones que se van dando en cada momento, incluyendo los pensamientos y las acciones conscientes, basadas en experiencias pasadas. La tesis del Trasfondo tiene que volverse a escribir otra vez eliminando la presuposición de la mente como una colección, un inventario, de fenómenos mentales, puesto que la única realidad ocurrente de los estados mentales en tanto que mentales es la conciencia.

La creencia en una realidad ocurrente que consta de estados mentales inconscientes, y que es distinta de las capacidades de Trasfondo, es una ilusión basada en gran medida en la gramática de nuestro lenguaje. Incluso cuando Jones está dormido, decimos que cree que Bush es presidente y que conoce las reglas de la gramática francesa. Así pensamos que allí dentro en su cerebro, durmiendo también, están su creencia de que Bush es presidente y su conocimiento del francés. Pero de hecho, todo lo que su cerebro contiene es un conjunto de estructuras neuronales, cuyo funcionamiento nos es bastante desconocido en la actualidad, que le capacitan a pensar y a actuar cuando llega el caso. Entre otras cosas, le capacitan para pensar que Bush es presidente y para hablar francés.

La mejor manera de pensar en estos asuntos es esta: en mi cerebro hay una enorme y compleja masa de neuronas incrustadas en las células gliales. Algunas veces la conducta de los elementos de esta masa compleja causa estados conscientes, incluyendo aquellos estados conscientes que son parte de las acciones humanas. Los estados conscientes tienen todo el color y la variedad que constituye nuestra vida de vigilia. Pero en el nivel de lo mental estos son todos los hechos. Lo que sucede en el cerebro, que es distinto de la conciencia, tiene una realidad ocurrente que es neurofisiológica más bien que psicológica. Cuando hablamos de estados inconscientes, estamos hablando de las capacidades del cerebro para generar conciencia. Además, algunas capacidades del cerebro no generan conciencia, sino que más bien funcionan para fijar la aplicación de los estados conscientes. Me capacitan para pasear, correr, escribir, hablar, etc.

Dado este cuadro, ¿cómo damos cuenta de todas aquellas intuiciones que nos llevaron a la tesis original del Trasfondo y a la distinción entre Trasfondo y Red? De acuerdo con la explicación que he dado en el capítulo anterior, cuando describimos un hombre en tanto que te-

niendo una creencia inconsciente, estamos describiendo una neurofisiología ocurrente en términos de su capacidad disposicional para causar pensamientos y conducta conscientes. Pero si esto es correcto, entonces parece seguirse que la Red de intencionalidad inconsciente es parte del Trasfondo. La ontología ocurrente de aquellas partes de la Red que son inconscientes es la de una capacidad neurofisiológica, pero el Trasfondo consta enteramente de tales capacidades.

La cuestión de cómo distinguir entre Red y Trasfondo desaparece, porque la Red es aquella parte del Trasfondo que describimos en términos de su capacidad para causar intencionalidad consciente. Pero todavía no estamos fuera del cenagal, pues nos queda la cuestión siguiente: ¿en qué se ha de convertir la tesis de que la intencionalidad funciona respecto de un conjunto de capacidades no intencionales? ¿Por qué la capacidad de generar la pregunta de que Bush es presidente ha de tratarse de manera diferente que, por ejemplo, la capacidad de generar la creencia de que los objetos son sólidos? ¿Y hemos de hacer una distinción entre el funcionamiento de la intencionalidad inconsciente y las capacidades no intencionales? Me parece que hemos cambiado el problema de distinguir entre Red y Trasfondo por el problema de distinguir lo intencional de lo no intencional dentro de las capacidades de Trasfondo.

Así pues, necesitamos hacer algunas distinciones más:

1. Necesitamos distinguir entre lo que está en el centro de nuestra atención consciente de las condiciones límite, periféricas, y de situación de nuestras experiencias conscientes, tal como se describen en el capítulo 6. En algún sentido esta es una distinción de prefondo-trasfondo, pero no nos interesa ahora.
2. Necesitamos distinguir dentro de los fenómenos mentales la forma representacional de la no representacional. Puesto que la intencionalidad se define en términos de representación, ¿cuál es el papel, si es que hay alguno, de lo no representacional en el funcionamiento de la intencionalidad?
3. Necesitamos distinguir capacidades de sus manifestaciones. Una de nuestras preguntas es: ¿cuáles de las capacidades del cerebro deberían pensarse como capacidades de Trasfondo?
4. Necesitamos distinguir aquello en lo que nos interesamos efectivamente de aquello que damos por sentado.

Estas distinciones se entrecruzan. A la luz de estas distinciones, y bajo el supuesto de que hemos abandonado la concepción de la mente como inventario, me parece que deberíamos volver a enunciar la hipótesis del Trasfondo como sigue:

Toda la intencionalidad consciente —todo pensamiento, percepción, comprensión, etc.— determina condiciones de satisfacción sólo relativamente a un conjunto de capacidades que no son y no pueden ser parte de ese mismo estado consciente. El contenido efectivo por sí mismo es insuficiente para determinar las condiciones de satisfacción.

De la intuición original de que los estados intencionales requieren un Trasfondo no intencional queda esto: incluso si se hacen explícitos todos los contenidos de la mente como un conjunto de reglas, pensamientos, creencias, etc., conscientes, aún se requiere un conjunto de capacidades de Trasfondo para su interpretación. Se ha perdido esto: no hay realidad ocurrente alguna en una Red inconsciente de intencionalidad, una Red que apoya holísticamente a todos sus miembros, pero que requiere un apoyo adicional del Trasfondo. En lugar de decir: «Para tener una creencia se tienen que tener muchas otras creencias», se debería decir: «Para tener un pensamiento consciente, se tiene que tener la capacidad de generar muchos otros pensamientos conscientes. Y esos pensamientos conscientes requieren todos ellos capacidades adicionales para su aplicación».

Ahora bien, dentro del conjunto de capacidades habrá algunas que se han adquirido en forma de reglas, hechos, etc., conscientemente aprendidos. Por ejemplo, a mí se me han enseñado las reglas del béisbol, que en Estados Unidos conducimos por la derecha, y el hecho de que George Washington fue el primer presidente. No se me enseñó regla alguna para andar, ni tampoco que los objetos son sólidos. La intuición original de que hay una distinción Red y Trasfondo se deriva de este hecho. Alguna de las capacidades que uno tiene nos capacitan para formular y aplicar reglas, principios, creencias, etc., en las realizaciones conscientes que uno lleva a cabo. Pero necesitan todavía capacidades de Trasfondo para su aplicación.

Si se empieza a pensar sobre la solidez de los objetos, entonces uno puede formarse una creencia consciente de que los objetos son sólidos. La creencia en la solidez de los objetos se convierte entonces en una creencia como cualquier otra, sólo que mucho más general.

De nuestras cinco tesis originales, tenemos ahora la siguiente lista revisada:

1. Los estados intencionales no funcionan autónomamente. No determinan sus condiciones de satisfacción independientemente.
2. Cada estado intencional requiere para su funcionamiento un conjunto de capacidades de Trasfondo. Las condiciones de satisfacción se determinan sólo relativamente a esas capacidades.
3. Entre esas capacidades estarán algunas que son capaces de generar otros estados conscientes. A estas otras se aplican las condiciones 1 y 2.
4. El mismo *tipo* de contenido intencional puede determinar diferentes condiciones de satisfacción cuando se manifiesta en diferentes instancias conscientes, de manera relativa a diferentes capacidades de Trasfondo, y relativamente a algunos Trasfondos no determina ninguna.

IV. MALAS COMPRENSIONES DEL TRASFONDO

Hay diversos modos de entender de mala manera la significación de la hipótesis del Trasfondo y quiero ahora eliminarlas. En primer lugar, muchos filósofos que son conscientes del Trasfondo están extremadamente desconcertados por él. Súbitamente les parece que el significado, la intencionalidad, la racionalidad, etc., se encuentran de alguna manera amenazados si su aplicación depende de hechos culturales y biológicos, cuya existencia es contingente, acerca de los seres humanos. Hay un cierto pánico que le sobreviene a un cierto tipo de sensibilidad filosófica cuando se reconoce que el proyecto de fundamentar la intencionalidad y la racionalidad en algunos cimientos puros, en algún conjunto de verdades necesarias e indudables, está, en principio, equivocado. Incluso les parece a algunas personas que es imposible tener una teoría del Trasfondo, puesto que el Trasfondo es la precondición de toda teoría, y en algunos casos extremos parece incluso como si cualquier teoría fuese imposible, puesto que la teoría depende de lo que parecen ser arenas movedizas de presuposiciones injustificables.

Contra este punto de vista, quiero decir que el descubrimiento del Trasfondo muestra solamente que una cierta concepción filosófica estaba equivocada. No amenaza ningún aspecto de nuestra vida diaria, incluyendo nuestra vida teórica diaria. Esto es, no muestra que el significado o la intencionalidad sean inestables o indeterminados, que nunca

nos podemos hacer entender, que la comunicación es imposible o está amenazada; muestra meramente que todo esto funciona respecto de un conjunto de capacidades y prácticas de Trasfondo que existen contingentemente. Además, la tesis del Trasfondo no muestra que la teorización es imposible; por el contrario, el Trasfondo mismo me parece un territorio excelente para la teorización, como espero que quede ilustrado por este capítulo.

Es también importante señalar que el Trasfondo no tiene implicaciones metafísicas, puesto que es un rasgo de nuestras *representaciones* de la realidad, y no un rasgo de la *realidad* representada. Algunos encuentran que es tentador pensar que, de acuerdo con la hipótesis del Trasfondo, la realidad misma se convierte, de uno u otro modo, en algo relativo al Trasfondo, y que, consecuentemente, debe seguirse algún género de relativismo o idealismo. Pero esto es un error. Al mundo real no le importa nada cómo lo representemos, y aunque nuestro sistema de representación exija un conjunto de capacidades no representacionales para funcionar, la realidad para cuya representación se usa ese sistema no depende en sí misma de esas capacidades ni de ninguna otra cosa. Dicho brevemente: el Trasfondo no amenaza nuestra convicción acerca del realismo externo, o la concepción de la verdad como correspondencia, o la posibilidad de comunicación clara, o la posibilidad de la lógica. Sin embargo, coloca a todos estos fenómenos bajo una luz diferente, puesto que no pueden proporcionar justificaciones trascendentales de nuestro discurso. Más bien, nuestra aceptación de ellos es una presuposición de Trasfondo del discurso.

Una mala comprensión del Trasfondo, particularmente importante en teorías de la interpretación textual, es la suposición errónea de que toda comprensión tiene que incluir algún acto de interpretación. Del hecho de que siempre que uno entiende algo lo entiende de una manera y no de otra, y del hecho de que son siempre posibles interpretaciones alternativas, simplemente no se sigue que en todo discurso uno esta siempre tomando parte en constantes «actos de interpretación». La compresión inmediata, normal, instantánea por parte de alguien de las emisiones es siempre posible sólo de manera relativa a un Trasfondo, pero no se infiere de este hecho que haya algún paso lógico separado, algún *acto* separado de interpretación que esté involucrado en la comprensión normal. Se comete un error similar en aquellas teorías de la cognición que afirman que tenemos que haber hecho siempre una inferencia si, cuando miramos un lado de un árbol, sabemos que ese árbol

tiene una parte posterior. Por el contrario, lo que hacemos es simplemente ver un árbol como un árbol real. Se podría, desde luego, dado un Trasfondo diferente, interpretar la propia percepción de manera diferente (por ejemplo, verlo como un estado bidimensional de una propiedad de árbol), pero del hecho de que uno tenga abiertas interpretaciones alternativas, no se sigue ni que las percepciones ordinarias involucren siempre un acto de interpretación ni que se dé algún paso inferencial, en tanto que proceso mental temporal efectivo, mediante el que se infieren datos no percibidos de datos percibidos.

El Trasfondo no es, quiero subrayarlo, un sistema de reglas. Esto, me parece a mí, era el punto débil de la noción de Foucault (1972) de la formación discursiva y la primera discusión de Bourdieu de una práctica en *Outline of a Theory of Practice* (1977). Ambos pensaban que las reglas eran esenciales a las clases de fenómenos que estoy tratando. Pero es importante ver que las reglas sólo tienen aplicaciones de manera relativa a las capacidades de Trasfondo. Las reglas no se autointerpretan y, en consecuencia, requieren un Trasfondo para funcionar; no son en sí ni explicativas ni constitutivas del Trasfondo.

A la luz de estas consideraciones, parece algunas veces como si el Trasfondo no pudiera representarse o hacerse completamente explícito. Pero esta formulación contiene ya un error. Cuando decimos esto tenemos ya un cierto modelo de representación y explicitud. La dificultad consiste en que el modelo es, simplemente, inaplicable al Trasfondo. *Desde luego*, el Trasfondo puede representarse. Aquí está: «el Trasfondo». Esta expresión representa el Trasfondo y, desde luego, el Trasfondo se puede hacer «completamente explícito» usando la misma expresión —o escribiendo un libro sobre el Trasfondo.

El asunto es que tenemos un modelo de explicitud para la representación de los fenómenos mentales que consiste en proporcionar oraciones que tengan el mismo contenido intencional que los estados representados. Puedo hacer completamente explícita la creencia de que el agua es húmeda diciendo, por ejemplo, que es la creencia de que el agua es húmeda. Pero dado que el Trasfondo no tiene ningún contenido intencional en este sentido, no podemos representarlo como si consistiese en un conjunto de contenidos intencionales. Esto no significa que no podamos describir el Trasfondo, o que su funcionamiento sea inanalizable, o cualquier otra cosa de este tipo. Son precisamente los comienzos de un análisis del Trasfondo lo que estoy intentando proporcionar.

V. RASGOS ADICIONALES DEL TRASFONDO

¿Podemos hacer una geografía del Trasfondo? ¿Podemos hacer una taxonomía de sus componentes? Bien, cualquier taxonomía exige principios de taxonomización. Hasta que no tengamos una noción clara de cómo funciona el Trasfondo, no seremos capaces de construir una taxonomía adecuada. Sin embargo, podemos empezar intuitivamente. En *Intentionality* (Searle, 1983) argumenté que necesitábamos, al menos, las siguientes distinciones: una distinción entre aquellos rasgos del Trasfondo que son comunes a todos los seres humanos y aquellos rasgos que tienen que ver con prácticas locales o culturales. Pongo en oposición estos dos grupos bajo el rótulo de «Trasfondo profundo» *versus* «prácticas locales». Las diferencias en Trasfondos locales hacen difícil la traducción de un lenguaje a otro; el que el Trasfondo profundo sea común es lo que la hace posible. Si se lee en Proust la descripción de una cena en casa de Guermantes, seguramente que se encontrará que alguno de los rasgos de la descripción son enigmáticos. Esto tiene que ver con diferencias en las prácticas culturales locales. Pero hay ciertas cosas que se pueden dar por sentadas. Por ejemplo, los participantes no comen llenándose las orejas de comida. Esto es un asunto del Trasfondo profundo. Hice también una distinción entre saber cómo hacer cosas y saber cómo son las cosas. Dicho de manera aproximada, esto intentaba capturar nuestra distinción tradicional entre lo práctico y lo teórico. Desde luego, tanto la razón práctica como la teórica dependen del Trasfondo y, por lo tanto, el Trasfondo mismo no es ni práctico ni teórico. Pero aún necesitamos hacer esta distinción. Un ejemplo de cómo hacer cosas es cómo andar. Un ejemplo de cómo son las cosas tendría que ver con la permanencia y estabilidad de los objetos que vemos en torno nuestro. Es obvio, sin embargo, que estas dos cosas están estrechamente relacionadas, puesto que no puede saberse cómo hacer cosas sin dar por sentado cómo son las cosas. No puedo, por ejemplo, «saber cómo» cortar madera sin dar por sentado que las hachas hechas de mantequilla no funcionarían y que las hachas hechas de agua no son hachas en absoluto.

He aquí algunas leyes de operación del Trasfondo. Algunas de ellas son:

1. En general, *no hay acción sin percepción, y no hay percepción sin acción.*

2. *La intencionalidad ocurre en un flujo coordinado de acción y percepción, y el Trasfondo es la condición de posibilidad de las formas*

tomadas por el flujo. Piensa en un segmento normal de tu vida de vigilia: estás tomando una comida, dando un paseo por el parque, haciendo el amor, o yendo a trabajar en el coche. En cada caso, la condición de posibilidad de la realización es una competencia de Trasfondo subyacente. El Trasfondo no sólo configura la aplicación del contenido intencional —lo que cuenta, por ejemplo, como «ir en el coche al trabajo»—, sino que la existencia del contenido intencional exige en primer lugar las capacidades de Trasfondo; sin un aparato complejo no se puede tener en modo alguno la intencionalidad involucrada en, por ejemplo, «ir en coche al trabajo».

3. *La intencionalidad tiende a elevarse al nivel de la capacidad de Trasfondo*. Así, por ejemplo, el esquiador principiante puede necesitar una intención de poner el peso en el esquí que va hacia abajo, un esquiador intermedio tiene la destreza que le capacita para tener la intención «vuelta a la derecha», y un esquiador realmente experto puede simplemente tener la intención «esquiar esta ladera». En una competición de esquí, por ejemplo, los entrenadores intentarán crear un nivel de intencionalidad que sea esencial para ganar la carrera, pero que presupone un enorme apuntalamiento por parte de las capacidades de Trasfondo. Así pues, el entrenador puede darle las siguientes intrucciones al esquiador: «Colócate cerca de las entradas en la manga, toma la entrada roja antes de empezar el descenso». De forma similar, cuando estoy hablando castellano, no tengo la intención de hacer concordar nombres en singular con verbos en singular o nombres en plural con verbos en plural; simplemente hablo.

4. Aunque la intencionalidad suba al nivel de la capacidad de Trasfondo, *alcanza la capacidad hasta el fondo*. Esta es otra manera de decir que todas las acciones subsidiarias voluntarias que se realizan dentro del alcance de una acción intencional de nivel superior son, a pesar de todo, intencionales. Así, por ejemplo, aunque no exijo una intención separada para mover mis brazos y mis piernas cuando esquío, o para mover mi boca cuando hablo, sin embargo todos esos movimientos se llevan a cabo intencionalmente.

Lo mismo sucede con la percepción. Normalmente no veo al nivel de manchas coloreadas; veo una ranchera Chevrolet con un parachoques delantero oxidado, o veo un cuadro de Vermeer con una mujer que está junto a una ventana, leyendo una carta, mientras que la luz se derrama por su vestido, la carta y la mesa. Pero obsérvese que en esos casos, aunque la intencionalidad de mi percepción suba al nivel de mi ca-

pacidad de Trasfondo (mi capacidad de reconocer rancheras Chevrolet, Vermeers, etc.), sin embargo los componentes de nivel inferior son también parte del contenido intencional; veo el azul de la ranchera y el marrón de la mesa.

5. *El Trasfondo sólo se manifiesta cuando hay contenido intencional.* Aunque el Trasfondo no es, en sí mismo, intencional, cualquier manifestación del Trasfondo, ya sea en la acción, en la percepción, etc., tiene que entrar en juego siempre que hay alguna intencionalidad, consciente o inconsciente. «El Trasfondo» no nombra una secuencia de eventos que puede simplemente ocurrir; más bien el Trasfondo consiste en capacidades mentales, disposiciones, posturas, modos de comportarse, saber cómo, *savoir faire*, etc., todos los cuales pueden sólo manifestarse cuando hay algún fenómeno intencional, tal como una acción intencional, una percepción, un pensamiento, etc.

9. LA CRÍTICA DE LA RAZÓN COGNITIVA

I. INTRODUCCIÓN: LOS MOVEDIZOS CIMIENTOS DE LA CIENCIA COGNITIVA

Durante una década, realmente desde los comienzos de la disciplina, he sido un «científico cognitivo» practicante. En este período he visto mucho trabajo y progreso valiosos en este campo. Sin embargo, como disciplina, la ciencia cognitiva adolece del defecto de que varias de sus más queridas suposiciones básicas son erróneas. Es posible realizar un buen trabajo a partir de supuestos falsos, pero es más difícil de lo que debería ser. En este capítulo quiero exponer y refutar alguna de estas suposiciones falsas. Derivan del modelo de errores que describí en los capítulos 1 y 2.

No todo el mundo está de acuerdo en ciencia cognitiva sobre los principios básicos, pero hay ciertos rasgos generales de la corriente principal que merecen ser enunciados separadamente. Si yo fuera un científico cognitivo adherido a la corriente principal, esto es lo que diría:

> Ni el estudio del cerebro como tal ni el estudio de la conciencia como tal tiene mucho interés ni importancia para la ciencia cognitiva. Los mecanismos cognitivos que estudiamos están, efectivamente, implementados en el cerebro, y algunos de ellos encuentran una expresión superficial en la conciencia, pero nuestro interés reside en el nivel intermedio donde los procesos cognitivos efectivos son inaccesibles a la conciencia. Aunque estén implementados de hecho en el cerebro, podrían haber estado implementados en un número indefinido de sistemas de *hardware*. Los cerebros están ahí, pero no son esenciales. Los procesos que explican la cognición son inconscientes no sólo de hecho, sino en principio. Por ejemplo, las reglas de Chomsky de la gramática universal

(1986), o las reglas de Marr de la visión (1982), o el lenguaje del pensamiento de Fodor (1975) no son la clase de fenómenos que podrían llegar a ser conscientes. Además, esos procesos son todos ellos computacionales. La suposición básica que está detrás de la ciencia cognitiva es que el cerebro es un ordenador [*computer*] y los procesos mentales son computacionales. Por esa razón muchos de nosotros pensamos que la inteligencia artificial (IA) es el corazón de la ciencia cognitiva. Hay alguna disputa entre nosotros sobre si el ordenador es o no un ordenador digital de la vieja variedad de von Neumann, o si es una máquina conexionista. Algunos de nosotros, de hecho, tenemos nuestra propia opinión sobre este asunto porque pensamos que los procesos en serie del cerebro se implementan por medio de un sistema conexionista paralelo (por ejemplo, Hobbs, 1990). Pero casi todos estamos de acuerdo en lo siguiente: los procesos mentales cognitivos son inconscientes; son, en su mayor parte, inconscientes en principio, y son computacionales.

No estoy de acuerdo con ninguna de las afirmaciones substantivas que se hacen en el parráfo anterior, y he criticado ya algunas de ellas en capítulos anteriores, notablemente la afirmación de que hay estados mentales que son profundamente inconscientes. El objetivo principal de este capítulo es criticar ciertos aspectos de la afirmación computacional.

Pienso que servirá de ayuda al explicar qué hace que el programa de investigación me parezca tan implausible que vinculemos la cuestión a un ejemplo concreto: en IA se han hecho grandes proclamas a favor de los programas ejecutados en SOAR.[1] Estrictamente hablando, SOAR es un tipo de arquitectura de ordenador, no un programa, pero los programas implementados en SOAR se consideran como ejemplos prometedores de IA. Uno de ellos está incorporado en un robot que puede mover bloques si se le ordena. Así, por ejemplo, el robot responderá adecuadamente a la orden: «Selecciona un bloque con forma de cubo y muévelo tres espacios a la izquierda». Para hacer esto, el robot tiene tanto unos sensores ópticos como unos brazos, y el sistema funciona porque implementa un conjunto de manipulaciones formales de símbolos que están conectadas con transductores que reciben *inputs* de los sensores ópticos y envían *outputs* a los mecanismos motores. Pero mi problema es este: ¿qué tiene que ver todo esto con la conducta hu-

1. SOAR es un sistema desarrollado por Alan Newell y sus colegas en la Universidad de Carnegie Mellon. El nombre es un acrónimo de *State, Operator, And Result.* Para una exposición, véase Waldrop (1988).

mana efectiva? Conocemos, por ejemplo, muchos de los detalles sobre cómo un ser humano hace esto en la vida real. En primer lugar, tiene que ser consciente. Además tiene que *oír* y *entender* la orden. Tiene que *ver conscientemente* los bloques, tiene que *decidir* llevar a cabo la orden y, a continuación, tiene que realizar la *acción intencional voluntaria consciente* de mover los bloques. Obsérvese que todas estas afirmaciones apoyan contrafácticos; por ejemplo: si no hay conciencia, no hay movimiento de bloques. También sabemos que toda esta maraña mental está causada por, y realizada en, la neurofisiología. Así pues, antes de que hayamos empezado con el modelo del ordenador sabemos que hay dos conjuntos de niveles: niveles mentales, muchos de ellos conscientes, y niveles neurofisiológicos.

Ahora bien, ¿dónde están las manipulaciones formales de símbolos que se supone que encajan en este cuadro? Es esta una cuestión fundamental por lo que respecta a los cimientos de la ciencia cognitiva, pero uno se asombra de la poca atención que se le presta. La cuestión absolutamente crucial para cualquier modelo de ordenador es «¿Cómo se relaciona *exactamente* el modelo con la realidad que ha de ser modelada?». Pero a menos que uno lea a críticos escépticos como el autor de este libro, no se encontrarán demasiadas discusiones de este problema. La respuesta general, que se supone que se evade de la exigencia de respuestas específicas más detalladas, es que entre el nivel de intencionalidad humana (lo que Newell [1982] llama «el nivel de conocimiento») y los diversos niveles neurofisiológicos, hay un nivel intermedio de manipulación de símbolos formales. Ahora bien, nuestra cuestión es: empíricamente hablando, ¿qué podría significar esto?

Si uno lee libros sobre el cerebro (por ejemplo, Shepherd, 1983, o Bloom y Lazerson, 1988), se obtiene una cierta visión de lo que ocurre en el cerebro. Si se pasa a libros sobre computación (por ejemplo, Boolos y Jeffrey, 1989), se obtiene una visión de la estructura lógica de la teoría de la computación. Si pasamos a su vez a libros sobre ciencia cognitiva (por ejemplo, Pylyshyn, 1984) nos encontraremos que nos dicen que lo que los libros sobre el cerebro describen es lo mismo que están describiendo los libros sobre computación. Filosóficamente hablando, esto no me huele nada bien y he aprendido, al menos al comienzo de una investigación, a seguir mi sentido del olfato.

II. IA FUERTE, IA DÉBIL Y COGNITIVISMO

La idea básica del modelo del ordenador de la mente es que la mente es el programa y el cerebro es el *hardware* de un sistema computacional. Un eslogan que se ve a menudo es el siguiente: «La mente es al cerebro lo que el programa es al *hardware*».[2]

Comencemos la investigación de este eslogan distinguiendo tres cuestiones:

1. ¿Es el cerebro un ordenador digital?
2. ¿Es la mente un programa de ordenador?
3. ¿Pueden simularse las operaciones del cerebro en un ordenador digital?

En este capítulo examinaré 1, pero no 2 o 3. En escritos anteriores (Searle, 1980a, 1980b, y 1984b), di una respuesta negativa a 2. Puesto que los programas se definen en términos puramente formales o sintácticos, y puesto que las mentes tienen un contenido mental intrínseco, se sigue inmediatamente que el programa por sí mismo no puede constituir una mente. La sintaxis formal del programa no garantiza por sí misma la presencia de contenidos mentales. Mostré esto hace una década en el argumento de la habitación china (Searle, 1980a). Un ordenador, yo por ejemplo, podría dar todos los pasos del programa para una capacidad mental, por ejemplo: para la capacidad mental de entender chino, sin entender una sola palabra de chino. El argumento descansa en la simple verdad lógica de que la sintaxis no es lo mismo que, ni es por sí misma suficiente para, la semántica. Así pues, la respuesta a la segunda pregunta es, demostrablemente, «No».

La respuesta a 3 me parece, de manera igualmente demostrable, que es «Sí», al menos de acuerdo con una interpretación natural. Esto es: naturalmente interpretada, la pregunta significa: ¿hay alguna descripción del cerebro tal que, bajo esta descripción, se podría hacer una simulación computacional de las operaciones del cerebro? Pero dada la tesis de

2. Este punto de vista se anuncia y se defiende en un gran número de libros y artículos, muchos de los cuales parecen tener el mismo título, por ejemplo, *Computers and Thought* (Feigenbaum y Feldman, eds., 1963), *Computers and Thought* (Sharples *et al.*, 1988), *The Computer and the Mind* (Johnson-Laird, 1988), *Computation and Cognition* (Pylyshyn, 1984), «The Computer Model of the Mind» (Block, 1990), y por supuesto, «Computing Machinery and Intelligence» (Turing, 1950).

Church de que cualquier cosa a la que pueda darse una caracterización suficientemente precisa en términos de un conjunto de pasos puede simularse en un ordenador digital, se sigue trivialmente que la cuestión tiene una respuesta afirmativa. Las operaciones del cerebro se pueden simular en un ordenador digital en el mismo sentido en que se pueden simular los sistemas climáticos, la conducta de la bolsa de Nueva York, o el modelo de los vuelos de las compañías aéreas sobre Latinoamérica. Así pues, nuestra cuestión no es «¿Es la mente un programa?». La respuesta a esto es «No». Ni tampoco es «¿Puede simularse el cerebro?». La respuesta a esto es «Sí». La cuestión es «¿Es el cerebro un ordenador digital?». Y para los propósitos de esta discusión considero esta pregunta como equivalente a: «¿Son los procesos cerebrales computacionales?».

Podría pensarse que esta pregunta perdería gran parte de su interés si la cuestión 2 recibe una respuesta negativa. Esto es: podría suponerse que a menos que la mente sea un programa, la cuestión de si el cerebro es un ordenador carece de interés. Pero este no es realmente el caso. Incluso para aquellos que están de acuerdo en que los programas no son por sí mismos constitutivos de los fenómenos mentales, queda todavía una cuestión importante: aceptemos que en la mente hay algo más que las operaciones sintácticas del ordenador digital; sin embargo, podría ser el caso que los estados mentales sean *al menos* estados computacionales, y los procesos mentales son procesos computacionales que operan sobre la estructura formal de esos estados mentales. Esta es, de hecho, la posición que me parece que adopta un número importante de gente.

No estoy diciendo que este punto de vista esté completamente claro, pero la idea es algo parecido a lo siguiente: en algún nivel de descripción, los procesos cerebrales son sintácticos; hay, por así decirlo, «oraciones en la cabeza». No es necesario que sean oraciones del castellano o del chino, sino que quizás pueden ser oraciones del «lenguaje del pensamiento» (Fodor, 1975). Ahora bien, al igual que cualesquiera oraciones, éstas tienen una estructura sintáctica y una semántica o significado, y el problema de la sintaxis puede separarse del problema de la semántica. El problema de la semántica es: ¿cómo obtienen esas oraciones que están en la cabeza sus significados? Pero tal cuestión puede discutirse independientemente de esta: ¿cómo funciona el cerebro al procesar esas oraciones? Una respuesta típica a esta última pregunta es: el cerebro funciona como un ordenador digital realizando operaciones computacionales sobre la estructura sintáctica de las oraciones que están en la cabeza.

En aras de mantener la terminología, llamo IA fuerte al punto de vista de acuerdo con el cual todo aquello en lo que consiste tener una mente es tener un programa, IA débil al punto de vista de que los procesos cerebrales (y los procesos mentales) pueden simularse computacionalmente, y cognitivismo al punto de vista de que el cerebro es un ordenador digital. Este capítulo trata sobre el cognitivismo.

III. La historia primigenia

Anteriormente he hecho una exposición preliminar de los supuestos de la corriente principal de la ciencia cognitiva, y quiero continuar ahora tratando de exponer, de la manera más firme que pueda, por qué el cognitivismo ha parecido intuitivamente atractivo. Hay toda una historia sobre la relación entre la inteligencia humana y la computación que se remonta al menos hasta el artículo clásico de Turing (1950) que creo que es el fundamento del punto de vista cognitivista. La llamaré la historia primigenia:

> Comenzamos con dos resultados de lógica matemática, la tesis de Church-Turing y el teorema de Turing. Para nuestros propósitos, la tesis de Church-Turing enuncia que para cualquier algoritmo hay alguna máquina de Turing que puede implementar el algoritmo. La tesis de Turing dice que hay una máquina universal de Turing que puede simular cualquier máquina de Turing. Ahora bien, si los ponemos juntos, obtenemos el resultado de que una máquina universal de Turing puede implementar un algoritmo cualquiera.

Pero ¿por qué este resultado es tan excitante? Bien, lo que hizo temblar y doblar el espinazo a toda una generación de jóvenes investigadores en inteligencia artificial fue el siguiente pensamiento: supongamos que el cerebro es una máquina universal de Turing.

Bien ¿hay alguna buena razón para suponer que el cerebro podría ser una máquina universal de Turing? Continuemos con la historia primigenia:

> Es claro que al menos algunas de las capacidades mentales humanas son algorítmicas. Por ejemplo, puedo hacer largas divisiones conscientemente recorriendo los pasos de un algoritmo para resolver problemas de divisiones largas. Además, es una consecuencia de la tesis de

> Church-Turing y del teorema de Turing que cualquier cosa que un humano puede hacer algorítmicamente puede hacerse con una máquina universal de Turing. Puedo implementar, por ejemplo, en un ordenador digital el mismo algoritmo que uso para divisiones largas. En tal caso, como describe Turing (1950), tanto yo, el ordenador humano, como el ordenador mecánico, estamos implementando el mismo algoritmo. Yo estoy haciéndolo conscientemente, el ordenador mecánico no conscientemente. Ahora parece razonable suponer que podría haber muchos otros procesos mentales que suceden en mi cerebro de manera no consciente y que son también computacionales. Y si esto es así, podríamos averiguar cómo funciona el cerebro simulando esos mismos procesos en un ordenador digital. Así como hemos logrado una simulación en un ordenador de los procesos de hacer divisiones largas, del mismo modo podríamos lograr una simulación en un ordenador de los procesos de comprender un lenguaje, percepción visual, categorización, etc.

«¿Pero qué sucede con la semántica? Despues de todo, los programas son puramente sintácticos.» Aquí entra en juego en la historia primigenia otro conjunto de resultados lógico-matemáticos:

> El desarrollo de la teoría de la demostración mostró que, dentro de ciertos límites bien conocidos, las relaciones semánticas entre proposiciones pueden reflejarse enteramente por medio de las relaciones sintácticas entre las oraciones que expresan esas proposiciones. Supongamos ahora que los contenidos mentales que están en la cabeza se expresan sintácticamente en la cabeza; entonces todo lo que necesitaríamos para dar cuenta de los procesos mentales serían procesos computacionales entre los elementos sintácticos que están en la cabeza. Si obtenemos la teoría de la demostración correcta, la semántica se cuidará de sí misma; y esto es lo que hacen los ordenadores: implementan la teoría de la demostración.[3]

Tenemos entonces un programa de investigación bien definido. Intentamos descubrir los programas que están implementados en el cere-

3. Todo este programa de investigación ha sido claramente resumido por Gabriel Segal (1991) de la manera siguiente: «La ciencia cognitiva contempla los procesos cognitivos como computaciones en el cerebro. Y la computación consiste en la manipulación de elementos sintácticos. El contenido de estos objetos sintácticos, si es que tienen alguno, es irrelevante para el modo en como se procesan. Así, parece que el contenido puede figurar en las explicaciones cognitivas sólo en tanto que las diferencias de contenido se reflejen en diferencias en la sintaxis del cerebro» (p. 463).

bro programando ordenadores para implementar los mismos programas. A su vez, hacemos esto intentando lograr que el ordenador mecánico se ajuste a las realizaciones del ordenador humano (esto es: que pase el test de Turing) y haciendo que los psicólogos busquen evidencia a favor de que los procesos internos son los mismos en los dos tipos de ordenador.

En lo que sigue me gustaría que el lector tuviera presente esta historia primigenia. Obsérvese especialmente el contraste de Turing entre la implementación consciente del programa por parte del ordenador humano y la implementación no consciente del programa, ya sea por el cerebro o por un ordenador mecánico. Téngase presente también la idea de que podríamos *descubrir* programas que se ejecutan en la naturaleza, los mismos programas que ponemos en nuestros ordenadores mecánicos.

Si se miran los libros y los artículos que apoyan el cognitivismo, se encuentran ciertas suposiciones comunes, a menudo sin enunciar, pero que, sin embargo, lo invaden todo.

En primer lugar, se supone a menudo que la única alternativa al punto de vista de que el cerebro es un ordenador digital es alguna forma de dualismo. He discutido las razones de este impulso en el capítulo 2. Retóricamente hablando, la idea es hacer que el lector piense que, a menos que acepte la idea de que el cerebro es algún género de ordenador, está comprometido con algunos puntos de vista extrañamente anticientíficos.

En segundo lugar, se supone que la cuestión de si los procesos cerebrales son computacionales es sólo una cuestión pura y simplemente empírica. Ha de establecerse por medio de investigación fáctica del mismo modo que cuestiones tales como si el corazón es un dispositivo de bombeo o si las hojas verdes sufren el proceso de fotosíntesis se establecieron en su momento como asuntos de hecho. No hay lugar aquí para realizar discriminaciones lógicas o llevar a cabo análisis conceptual, puesto que estamos hablando sobre asuntos de hecho, de hechos científicos concretos. De hecho, pienso que mucha gente que trabaja en este campo dudaría de que la cuestión que estoy planteando sea una cuestión filosófica apropiada. «¿Es realmente el cerebro un ordenador digital?» no es una cuestión filosófica con más derecho que «¿es realmente el acetilcoleno el neurotransmisor de las junturas neuromusculares?».

Incluso gente que no simpatiza con el cognitivismo, por ejemplo Penrose (1989) y Dreyfus (1972), parecen tratarlo como una posición

lisa y llanamente fáctica. No parecen preocuparse por la cuestión de qué tipo de afirmación podría ser aquella de la que dudan. Pero a mí me resulta problemática la pregunta: ¿qué clase de hecho sobre el cerebro constituiría su ser un ordenador?

En tercer lugar, otro rasgo estilístico de las publicaciones en este campo es la preocupación e incluso la falta de rigor con la que se glosan las cuestiones básicas. ¿Cuáles son exactamente los rasgos anatómicos y fisiológicos del cerebro que se discuten? ¿Qué es exactamente un ordenador digital? ¿Y cómo se supone que están conectadas las respuestas a estas dos preguntas? El procedimiento usual en esos libros y artículos es hacer un puñado de observaciones sobre ceros y unos, dar un resumen divulgativo de la tesis de Church-Turing y, a continuación, pasar a cosas más excitantes como los logros y los fallos del ordenador. Para mi sorpresa he encontrado, al leer estos libros y artículos, lo que parece ser un hiato filosófico particular. Por una parte, tenemos un conjunto muy elegante de resultados matemáticos que van desde el teorema de Turing a la teoría de funciones recursivas, pasando por el teorema de Church. Por otro lado, tenemos un conjunto impresionante de dispositivos electrónicos que usamos todos los días. Puesto que tenemos una matemática tan avanzada y una electrónica tan buena suponemos que alguien tiene que haber hecho de alguna manera el trabajo filosófico básico de conectar la matemática con la electrónica. Pero todo lo que puedo decir es que este no es el caso. Por el contrario, estamos en una situación peculiar donde hay poco acuerdo teórico entre los participantes en la discusión sobre cuestiones absolutamente fundamentales como ¿qué es exactamente un ordenador digital? ¿Qué es exactamente un símbolo? ¿Qué es exactamente un algoritmo? ¿Qué es exactamente un proceso computacional? ¿Bajo qué condiciones físicas están dos sistemas implementando exactamente el mismo programa?

IV. LA DEFINICIÓN DE COMPUTACIÓN

Puesto que no hay acuerdo universal sobre las cuestiones fundamentales, creo que lo mejor es volver a las fuentes, volver a las definiciones dadas por Alan Turing.

De acuerdo con Turing, una máquina de Turing puede llevar a cabo ciertas operaciones elementales: puede volver a escribir un 0 de su cin-

ta como un 1, puede volver a escribir un 1 de su cinta como un 0, puede mover la cinta un cuadrado a la izquierda, o puede mover la cinta un cuadrado a la derecha. Está controlada por un programa de instrucciones y cada instrucción especifica una condición y una acción que ha de llevarse a cabo si se satisface la condición.

Esta es la definición estándar de computación, pero tomada literalmente es, al menos, desorientadora. Si uno pone en marcha su ordenador, es muy poco probable que se encuentre con ceros y unos y ni siquiera con una cinta. Pero esto no importa realmente para la definición. Para averiguar si un objeto es realmente un ordenador digital resulta que no tenemos que buscar efectivamente ni ceros ni unos; más bien tenemos que buscar algo que pudiese ser *tratado como*, o que *contase como*, o que *pudiese ser usado para* funcionar como ceros y unos. Además, para que el asunto sea más problemático, resulta que esta máquina podría estar hecha de cualquier cosa. Como dice Johnson-Laird: «Podría estar hecha de ruedas dentadas y palancas igual que una calculadora antigua; podría estar constituida por un sistema hidráulico a través del que fluye el agua; podría estar hecha de transistores incrustados en un chip de silicio a través del cual fluye la corriente; incluso podría ser el cerebro. Cada una de esas máquinas usa un medio diferente para representar símbolos binarios. Las posiciones de las ruedas dentadas, la presencia o ausencia de agua, el voltaje y quizás los impulsos nerviosos» (Johnson-Laird, 1988, p. 39).

La mayor parte de la gente que escribe sobre este tema hace observaciones similares. Por ejemplo, Ned Block (1990) muestra cómo podemos tener pasos eléctricos donde los unos y los ceros se asignan a los niveles de voltaje de 4 y 7 voltios respectivamente. Así, podríamos pensar que deberíamos buscar los niveles de voltaje. Pero Block nos dice que 1 se asigna sólo «convencionalmente» a cierto nivel de voltaje. La situación se vuelve más problemática cuando nos informa a continuación de que no necesitamos en absoluto utilizar electricidad, sino que podemos usar un elaborado sistema de gatos, ratones y queso y hacer que nuestros pasos sean tales que el gato tirará de la cuerda y abrirá un paso que también podemos tratar como si fuera un 0 o un 1. El asunto crucial es, como Block insiste una y otra vez, «la irrelevancia de la realización física del *hardware* para la descripción computacional. Esos pasos funcionan de maneras diferentes pero, con todo, son computacionalmente equivalentes» (p. 260). Del mismo modo, Pylyshyn dice que una secuencia computacional podría ser realizada por «¡un grupo

de palomas entrenadas para picotear como una máquina de Turing!» (1984, p. 57).

Ahora bien, si estamos intentando tomar en serio la idea de que el cerebro es un ordenador digital, lo que obtenemos es el poco confortable resultado de que podríamos hacer un sistema prácticamente de cualquier cosa que haga precisamente lo que el cerebro hace. Computacionalmente hablando, se podría hacer un «cerebro» que funcionase como el tuyo y el mío a partir de componentes como gatos, ratones y queso, palancas, bombas de agua, palomas, o cualquier otra cosa siempre que los dos sistemas fuesen, en el sentido de Block, «computacionalmente equivalentes». Sólo se necesitaría una terrible cantidad de gatos, o palomas, o bombas de agua, o cualquier otra cosa. Los proponentes del cognitivismo informan de este resultado con total y no disimulada complacencia. Pero pienso que les debería causar preocupación, y voy a intentar mostrar que esta es sólo la punta de todo un iceberg de problemas.

V. Primera dificultad: la sintaxis no es intrínseca a la física

¿Por qué no están preocupados los defensores del computacionalismo por las implicaciones de la realización múltiple? La razón es que piensan que es típico de las explicaciones funcionales el que la misma función admite realizaciones múltiples. A este respecto, los ordenadores son como los carburadores y los termostatos. Lo mismo que los carburadores pueden hacerse de bronce o de acero, así los ordenadores pueden hacerse de un rango indefinido de materiales de *hardware*.

Pero hay una diferencia: las clases de los carburadores y de los termostatos se definen en términos de la producción de ciertos efectos *físicos*. Por esta razón, nadie te dice, por ejemplo, que se puedan hacer carburadores a base de palomas. Pero la clase de los ordenadores se define sintácticamente en términos de *asignaciones* de ceros y unos. La realizabilidad múltiple no es una consecuencia del hecho de que el mismo efecto físico pueda lograrse en diferentes substancias físicas, sino del hecho de que las propiedades relevantes son puramente sintácticas. La física es irrelevante excepto en el punto que admite asignaciones de ceros y unos y de transiciones de estado entre ellas.

Pero esto tiene dos consecuencias que podrían ser desastrosas:

1. El mismo principio que implica la realizabilidad múltiple parecería implicar la realizabilidad universal. Si la computación se define en términos de la asignación de sintaxis, entonces todo puede ser un ordenador digital, puesto que a cualquier objeto se le podrían hacer adscripciones sintácticas. Se podría describir cualquier cosa en términos de ceros y unos.
2. Peor aún, la sintaxis no es intrínseca a la física. La adscripción de propiedades sintácticas es siempre relativa a un agente u observador que trata como sintácticos ciertos fenómenos físicos.

Ahora bien, ¿por qué exactamente serían desastrosas estas consecuencias? Bien, queremos saber cómo funciona el cerebro, específicamente cómo produce fenómenos mentales. Y no sería una respuesta a esta pregunta el que se nos dijese que el cerebro es un ordenador digital en el sentido en que el estómago, el hígado, el corazón, el sistema solar y el estado de Kansas son todos ellos ordenadores digitales. El modelo que teníamos era que podríamos descubrir algún hecho sobre la operación del cerebro que mostraría que es un ordenador. Queríamos saber si no había algún sentido en el que los cerebros eran *intrínsecamente* ordenadores digitales del modo en el que las hojas verdes realizan intrínsecamente la fotosíntesis o los corazones bombean sangre intrínsecamente. No se trata de que asignemos arbitraria o «convencionalmente» la palabra «bombean» a los corazones o «fotosíntesis» a las hojas. Hay, efectivamente, un hecho objetivo. Y lo que estamos preguntando es: «¿hay en este sentido algún hecho objetivo respecto de los cerebros que los haga ordenadores digitales?». Y no constituye una respuesta a esta pregunta el que que se diga: sí, los cerebros son ordenadores digitales porque todo es un ordenador digital.

De acuerdo con la definición de computación de los manuales estándar, es difícil ver cómo pueden evitarse los resultados siguientes:

1. Para cualquier objeto hay una descripción de ese objeto tal que, bajo esa descripción, el objeto es un ordenador digital.
2. Para cualquier programa y para cualquier objeto suficientemente complejo, hay alguna descripción del objeto bajo la cual éste está implementando el programa. Así, por ejemplo, la pared que está detrás de

mí está implementando el programa *Wordstar*, puesto que hay algún modelo de movimientos moleculares que es isomórfico con la estructura formal del *Wordstar*. Pero si esa pared está implementando *Wordstar* entonces, si es una pared lo suficientemente grande, está implementando cualquier programa, incluyendo cualquier programa implementado en el cerebro.

Pienso que la principal razón por la que los proponentes no ven que la realización múltiple o universal es un problema es que no la ven como una consecuencia de un punto mucho más importante, a saber: que «sintaxis» no es el nombre de un rasgo físico, como masa o gravedad. Por el contrario, hablan de «motores sintácticos» e incluso de «motores semánticos» como si tal manera de hablar fuese semejante a aquella en la que se habla de motores de gasolina o motores diesel, como si pudiera ser un simple hecho objetivo que el cerebro, o cualquier otra cosa, sea un motor sintáctico.

No pienso que el problema de la realizabilidad universal sea serio. Pienso que es posible bloquear el resultado de la realizabilidad universal haciendo más estricta nuestra definición de computación. Ciertamente, deberíamos respetar el hecho de que los programadores y los ingenieros lo consideren como una peculiaridad de las definiciones originales de Turing y no como un rasgo real de la computación. Los trabajos no publicados de Brian Smith, Vinod Goel y John Batali sugieren que una definición más realista de la computación subrayaría rasgos tales como las relaciones causales entre estados de programa, programabilidad, y controlabilidad del mecanismo y situación en el mundo real. Todo esto produciría el resultado de que el modelo no es suficiente. Tiene que haber una estructura causal suficiente para garantizar contrafácticos. Pero estas restricciones adicionales sobre la noción de computación no sirven de ayuda en la presente discusión *porque el problema realmente profundo es que la sintaxis es esencialmente una noción relativa al observador. La realizabilidad múltiple de los procesos computacionalmente equivalentes en diferentes medios físicos no es sólo una señal de que los procesos son abstractos, sino de que no son en absoluto intrínsecos al sistema. Dependen de una interpretación desde fuera.* Buscamos algunos hechos objetivos que harían computacionales los procesos cerebrales; pero dado el modo en que hemos definido la computación, jamás puede haber tales hechos objetivos. No podemos, por un lado, decir que cualquier cosa es un ordenador digital

si podemos asignarle una sintaxis y suponer, por otro, que hay una cuestión fáctica intrínseca a su operación física que decide si un sistema natural tal como el cerebro es un ordenador digital.

Y si la palabra «sintaxis» parece problemática, puede enunciarse el mismo asunto sin hacer uso de ella. Esto es: alguien podría afirmar que las nociones de «sintaxis» y «símbolos» son sólo una manera de hablar y que aquello en lo que estamos realmente interesados es en la existencia de sistemas con fenómenos físicos discretos y en enunciar las transiciones entre ellos. De acuerdo con este punto de vista no necesitamos ningún 0 ni ningún 1; son sólo una abreviatura conveniente. Un estado físico de un sistema es un estado computacional sólo de manera relativa a la asignación a ese estado de algún rol, función o interpretación computacional. El mismo problema surge con 0 y 1, puesto que *nociones tales como computación, algoritmo y programa no nombran rasgos físicos intrínsecos de sistemas.* Los estados computacionales no se *descubren dentro de* la física, se *asignan a* la física.

Este es un argumento diferente del de la habitación china y debería haberlo visto hace diez años, pero no lo hice. El argumento de la habitación china mostraba que la semántica no es intrínseca a la sintaxis. Ahora estoy haciendo hincapié en un asunto separado y diferente: que la sintaxis no es intrínseca a la física. Para los propósitos del argumento original estaba suponiendo simplemente que la caracterización sintáctica del ordenador no era problemática. Pero esto es un error. No hay ninguna manera en que pueda descubrirse que algo es intrínsecamente un ordenador digital, puesto que su caracterización como ordenador digital es siempre relativa a un observador que asigna una interpretación sintáctica a los rasgos puramente físicos del sistema. Aplicado a la hipótesis del lenguaje del pensamiento, tiene la consecuencia de que la tesis es incoherente. No hay manera en que pueda descubrirse que hay, intrínsecamente, oraciones desconocidas que están en la cabeza, puesto que algo es una oración sólo relativamente a algún agente o usuario que la usa como oración. Aplicado generalmente al modelo computacional, la caracterización de un proceso como computacional es una caracterización de un sistema físico desde fuera; y la identificación del proceso como computacional no identifica un rasgo intrínseco de la física; es esencialmente una caracterización relativa al observador.

Este punto ha de entenderse de manera precisa. No estoy diciendo que haya límites *a priori* respecto de los patrones que podemos descubrir en la naturaleza. Sin duda, podríamos descubrir un patrón de

eventos en mi cerebro que fuera isomórfico con la implementación del programa de edición de textos de mi ordenador. Pero decir que algo está *funcionando como* un proceso computacional es decir algo más que está ocurriendo un patrón de eventos físicos. Requiere la asignación de una interpretación computacional por parte de un agente. Análogamente, podríamos descubrir objetos en la naturaleza que tuviesen la misma clase de contorno que las sillas y que, por lo tanto, pudiesen ser usadas como sillas; pero no podríamos descubrir en la naturaleza objetos que estuviesen funcionando como sillas, excepto de manera relativa a algunos agentes que las consideran o las usan como sillas.

Para entender completamente este argumento es esencial entender la distinción entre rasgos del mundo que son *intrínsecos* y rasgos que son *relativos al observador*. Las expresiones «masa», «atracción gravitatoria» y «molécula» nombran rasgos del mundo que son intrínsecos. Si todos los observadores y usuarios dejasen de existir, el mundo contendría aún masa, atracción gravitatoria y moléculas. Pero expresiones tales como «día precioso para ir a merendar al campo», «bañera» y «silla» no nombran rasgos intrínsecos de la realidad. Más bien, nombran objetos especificando algún rasgo que les ha sido asignado, algún rasgo que es relativo a observadores y usuarios. Si no hubiese habido jamás usuario u observador alguno, habría con todo montañas, moléculas, masas y atracción gravitatoria. Pero si no hubiese habido nunca ningún usuario u observador, no habría rasgos tales como ser un día precioso para ir a merendar al campo, o ser una silla o una bañera. La asignación de rasgos relativos al observador a rasgos intrínsecos del mundo no es arbitraria. Algunos rasgos intrínsecos del mundo facilitan su uso como, por ejemplo, sillas o bañeras. Pero el rasgo de ser una silla o una bañera o un día precioso para ir al campo a merendar es un rasgo que sólo existe de manera relativa a usuarios u observadores. Lo que quiero decir aquí, y que es la esencia de este argumento, es que de acuerdo con las definiciones estándar de computación, los rasgos computacionales son relativos al observador. No son intrínsecos. El argumento, tal como se ha enunciado hasta aquí, puede resumirse de la siguiente manera:

El objetivo de la ciencia natural es describir y caracterizar rasgos que son intrínsecos al mundo natural. Por sus propias definiciones de computación y cognición, no hay modo en que la ciencia cognitiva computacional pueda ser jamás una ciencia natural, puesto que la

computación no es un rasgo intrínseco del mundo. Se asigna de manera relativa a observadores.[4]

VI. Segunda dificultad: la falacia del homúnculo es endémica en el cognitivismo

Parece que ahora nos encontramos con un problema. La sintaxis no es parte de la física. Esto tiene como consecuencia el que si la computación se define sintácticamente, entonces nada es intrínsecamente un ordenador digital solamente en virtud de sus propiedades físicas. ¿Hay algún modo de escapar a esta dificultad? Sí, lo hay, un modo que se abraza de manera estándar en ciencia cognitiva, pero que equivale a pasar de Guatemala a Guatepeor. La mayor parte de las obras que he visto sobre teoría computacional de la mente cometen alguna variante de la falacia del homúnculo. La idea es siempre tratar el cerebro como si hubiese algún agente dentro de él que lo usase para realizar computaciones. Un caso típico es David Marr (1982), que describe la tarea de la visión como algo que tiene como origen una ordenación visual bidimensional en la retina y que produce como *output* del sistema visual una descripción tridimensional del mundo externo. La dificultad es: ¿quién lee la descripción? De hecho, a lo largo de todo el libro de Marr, y en otros libros estándar sobre el tema, parece como si tuviésemos que invocar un homúnculo que estuviese dentro del sistema para tratar sus operaciones como genuinamente computacionales.

Muchos autores piensan que la falacia del homúnculo no es realmente un problema porque, con Dennett (1978), tienen la sensación de que se puede «descargar» el homúnculo. La idea es: puesto que las operaciones computacionales del ordenador pueden analizarse en unidades progresivamente más simples, hasta que alcanzamos modelos simples de encendido-apagado, «sí-no», «1-0», parece que los homúnculos de nivel superior pueden descargarse en homúnculos más estúpidos, hasta que finalmente alcanzamos el nivel inferior de un simple apagado-encendido que no involucra homúnculos en absoluto. La idea,

4. Pylyshyn está muy cerca de aceptar precisamente esto cuando escribe: «La respuesta a la pregunta de qué computación se está realizando requiere la discusión de estados computacionales semánticamente interpretados» (1984, p. 58). Efectivamente. ¿Y quién está haciendo la interpretación?

dicho brevemente, es que la descomposición recursiva eliminará los homúnculos.

Me costó bastante tiempo averiguar qué quería decir esta gente, así que si alguien tiene los mismos problemas que yo tenía le voy a explicar con detalle un ejemplo. Supongamos que tenemos un ordenador que multiplica seis por ocho para obtener cuarenta y ocho. Y preguntamos: «¿cómo lo hace?». Bien, la respuesta podría ser que suma siete veces seis a sí mismo.[5] Pero si se pregunta: «¿cómo suma siete veces seis a sí mismo?», la respuesta podría ser que, en primer lugar, convierte todos los numerales en notación binaria y, en segundo lugar, aplica un algoritmo simple para operar en notación binaria hasta que, finalmente, alcanzamos el nivel inferior en el que las únicas instrucciones son de la forma: «Escribe un cero, borra un uno». Así, por ejemplo, en el nivel superior nuestro homúnculo inteligente dice: «Sé cómo multiplicar seis por ocho para obtener cuarenta y ocho». Pero en el nivel inferior inmediato es reemplazado por un homúnculo más estúpido que dice: «No sé como multiplicar, pero sé sumar». Detrás de él hay homúnculos más estúpidos que dicen: «No sabemos ni multiplicar ni sumar, pero sabemos cómo convertir la notación decimal en binaria». Detrás de éstos hay a su vez otros más estúpidos que dicen: «No sabemos nada de todo este asunto, pero sabemos cómo operar con símbolos binarios». En el nivel máximamente inferior hay toda una serie de homúnculos que lo único que dicen es «cero uno, cero uno». Todos los niveles superiores se reducen a este nivel inferior. Sólo existe realmente este nivel máximamente inferior; los niveles superiores son *como si* existieran.

Diversos autores (por ejemplo, Haugeland, 1981; Block, 1990) describen este rasgo cuando dicen que el sistema es un motor sintáctico que gobierna un motor semántico. Pero aún debemos hacer frente a la cuestión que teníamos delante con anterioridad: ¿qué hechos intrínsecos al sistema lo hacen sintáctico? ¿Qué hechos del nivel máximamente inferior, o de cualquiera de los otros niveles, hacen que esas operaciones sean sobre ceros y unos? *Sin un homúnculo que esté fuera de la descomposición recursiva, no tenemos ni siquiera una sintaxis para operar con ella.* El intento de eliminar la falacia del homúnculo me-

5. La gente dice algunas veces que debería haber sumado seis *ocho* veces a sí mismo. Pero esto es mala aritmética. Seis sumado ocho veces a sí mismo es cincuenta y cuatro, puesto que seis sumado cero veces a sí mismo es todavía seis. Es sorprendente la frecuencia con la que se comete este error.

diante descomposición recursiva falla, porque la única manera de obtener la sintaxis intrínseca a la física es poner un homúnculo en la física.

Hay un rasgo fascinante de todo esto. Los cognitivistas conceden de muy buen grado que los niveles de computación superior, por ejemplo, «multiplica seis por ocho», son relativos al observador; no hay realmente nada que corresponda directamente a la multiplicación; todo está en el ojo del homúnculo/observador. Pero quieren detener esta concesión en los niveles más bajos. El circuito electrónico, admiten, no multiplica realmente 6×8 como tal, sino que realmente manipula ceros y unos y esas manipulaciones, por así decirlo, equivalen a la multiplicación. Pero conceder que los niveles más elevados de computación no son intrínsecos a la física es conceder ya que los niveles inferiores tampoco lo son. Así pues, la falacia del homúnculo está todavía con nosotros.

Para el caso de los ordenadores reales que se compran en las tiendas, el problema del homúnculo no se presenta porque cada usuario es el homúnculo en cuestión. Pero si hemos de suponer que el cerebro es un ordenador digital, tenemos que hacer frente a la cuestión: «¿y quién es el usuario?». Las cuestiones típicas del homúnculo en ciencia cognitiva son del tipo siguiente: «¿cómo computa el sistema visual el contorno a partir del sombreado?»; «¿cómo computa la distancia del objeto a partir de la imagen retiniana?». Una pregunta paralela sería: «¿cómo computan los clavos la distancia que han de atravesar en la madera a partir del impacto del martillo y la densidad de la madera?». Y la respuesta es la misma en las dos clases de caso: si estamos hablando de cómo funciona el sistema intrínsecamente, ni los clavos ni los sistemas visuales computan nada. Podríamos describirlos computacionalmente como homúnculos externos, y esto es a menudo algo útil. Pero no entendemos el martillear suponiendo que los clavos están implementando intrínsecamente algo así como los algoritmos del martilleo, y no se entiende la visión suponiendo que el sistema está implementando, por ejemplo, el contorno a partir del algoritmo del sombreado.

VII. Tercera dificultad: la sintaxis no tiene poderes causales

Ciertas clases de explicaciones en ciencias naturales especifican mecanismos que funcionan causalmente en la producción de los fenómenos que han de explicarse. Esto es especialmente común en las cien-

cias biológicas. Piénsese en la teoría de la enfermedad causada por gérmenes, la explicación de la fotosíntesis, la teoría de los caracteres heredados del ADN, e incluso en la teoría darwiniana de la selección natural. En cada caso se especifica un mecanismo causal, y en cada caso la especificación da una explicación del *output* del mecanismo. Si uno vuelve atrás y le echa una ojeada a la historia primigenia parece claro que esta es la clase de explicación que promete el cognitivismo. Se supone que los mecanismos por los que el cerebro produce cognición son computacionales, y al especificar los programas hemos especificado las causas de la cognición. Parte de la belleza de este programa de investigación, que se subraya a menudo, es que no necesitamos conocer los detalles del funcionamiento del cerebro para explicar la cognición. Los procesos cerebrales proporcionan sólo el *hardware* en el que se implementan los programas cognitivos, pero es en el nivel del programa donde se dan las explicaciones cognitivas reales. De acuerdo con la explicación estándar, tal como la enuncia, por ejemplo, Newell (1982), hay tres niveles de explicación —*hardware*, programa e intencionalidad (Newell llama a este último nivel el nivel de conocimiento)— y la contribución especial de la ciencia cognitiva se hace en el nivel del programa.

Pero si lo que he dicho hasta ahora es correcto, entonces hay algo sospechoso en todo este proyecto. Yo solía creer que, en tanto que explicación causal, la teoría de los cognitivistas era por lo menos falsa, pero ahora tengo dificultades para formular una versión de ella que sea coherente incluso con el punto de vista de que podría ser una tesis empírica. La tesis es que hay toda una gran cantidad de símbolos que se están manipulando en el cerebro, ceros y unos que centellean de un lado a otro del cerebro a la velocidad de la luz y que son invisibles no sólo a simple vista, sino con el más potente microscopio electrónico, y que causan la cognición. Pero la dificultad es que los ceros y los unos como tales no tienen poderes causales puesto que ni siquiera existen excepto en los ojos del observador. El programa implementado no tiene poderes causales distintos del medio que lo implementa puesto que el programa no tiene existencia real, no tiene ontología más allá del medio que lo implementa. Físicamente hablando, no nay nada que sea un «nivel de programa» separado.

Se puede ver esto si volvemos a la historia primigenia y recordamos la diferencia entre el ordenador mecánico y el ordenador humano de Turing. En el ordenador humano de Turing hay realmente un nivel

de programa intrínseco al sistema, y está funcionando causalmente a este nivel para convertir *input* en *output*. Esto sucede así porque el ser humano está siguiendo conscientemente las reglas para llevar a cabo una cierta computación, y esto explica causalmente su ejecución. Pero cuando programamos el ordenador mecánico para realizar la misma computación, la asignación de una interpretación computacional es ahora relativa a nosotros, los homúnculos externos. No hay causación intencional intrínseca al sistema. El ordenador humano está siguiendo conscientemente reglas, y este hecho explica su conducta, pero el ordenador mecánico no está siguiendo literalmente regla alguna. Está diseñado para comportarse exactamente como si estuviera siguiendo reglas; así pues, para propósitos prácticos y comerciales, no importa que no esté siguiendo efectivamente regla alguna. No podría estar siguiendo reglas puesto que no tiene contenido intencional intrínseco al sistema que esté funcionando intencionalmente para producir la conducta. Ahora bien, el cognitivismo nos dice que el cerebro funciona como el ordenador comercial y que esto causa la cognición. Pero sin un homúnculo, tanto el ordenador comercial como el cerebro tienen únicamente modelos y los modelos no tienen poderes causales adicionales a los que tienen los medios que los implementan. De esta manera, parece que no hay ningún modo en el que el cognitivismo *pueda* dar una explicación causal de la cognición.

Mi punto de vista tiene, sin embargo, un problema. Cualquiera que trabaje con los ordenadores sabe incluso causalmente que a menudo damos de hecho explicaciones causales que apelan al programa. Por ejemplo, podemos decir que cuando pulso esta tecla obtengo tales y tales resultados porque la máquina está implementando este programa y no aquel; y esto parece una explicación causal ordinaria. De este modo, el problema es ¿cómo reconciliamos el hecho de que la sintaxis, como tal, no tenga poderes causales de ningún tipo con el hecho de que demos explicaciones causales que apelan a los programas? Y de manera más apremiante, ¿proporcionarían esas clases de explicación un modelo apropiado para el cognitivismo?, ¿salvarán el cognitivismo? ¿Podrían, por ejemplo, salvar la analogía con los termostatos señalando que así como la noción de «termostato» figura en las explicaciones causales independientemente de cualquier referencia a la física de su implementación, del mismo modo la noción de «programa» podría ser explicativa a la vez que igualmente independiente de la física?

Para explorar este problema, intentemos defender el cognitivismo

extendiendo la historia primigenia para mostrar cómo funcionan los procedimientos de investigación del cognitivista en la práctica de investigación efectiva. La idea, típicamente, es programar un ordenador comercial de modo que simule alguna capacidad cognitiva tal como la visión o el lenguaje. De este modo, si obtenemos una buena simulación, una simulación que nos dé al menos una equivalencia de Turing, establecemos la hipótesis de que el ordenador cerebral está ejecutando el mismo programa que el ordenador comercial, y para poner a prueba la hipótesis buscamos evidencia psicológica indirecta como, por ejemplo, los tiempos de reacción. Así pues, parece que podemos explicar causalmente la conducta del ordenador cerebral citando el programa en exactamente el mismo sentido en que podemos explicar la conducta del ordenador comercial. ¿Qué es lo que está mal aquí? ¿No suena esto de manera semejante a un programa de investigación científica perfectamente legítimo? Sabemos que la conversión de *input* a *output* que realiza el ordenador comercial se explica mediante un programa, y en el cerebro descubrimos el mismo programa; por consiguiente tenemos una explicación causal.

Hay dos cosas en este proyecto que deberían preocuparnos inmediatamente. En primer lugar, no deberíamos aceptar jamás este modo de explicación para cualquier función del cerebro de la que entendemos efectivamente cómo funciona en el nivel neurobiológico. En segundo lugar, no deberíamos aceptarla para otras clases de sistemas que podemos simular computacionalmente. Para ilustrar el primero de estos puntos considérese, por ejemplo, la famosa explicación de «What the Frog's Eye Tells to the Frog's Brain» (Lettvin *et al.,* 1959, en McCulloch, 1965). La explicación se da enteramente en términos de la anatomía y fisiología del sistema nervioso de la rana. Un pasaje típico, escogido al azar, reza del modo siguiente:

> *Detectores de contraste sostenidos*
>
> Un axón no mielinizado de este grupo no responde cuando la iluminación general se enciende o se apaga. Si el canto de un objeto se mueve dentro de su campo y se para, se descarga inmediatamente y continúa descargándose sin que importe cuál sea el contorno del canto o si el objeto es más pequeño o mayor que el campo receptivo (p. 239).

No he oído nunca a nadie decir que todo esto es sólo la implementación de *hardware* y que debería haberse averiguado qué programa es-

taba implementando la rana. No dudo de que se pueda hacer una simulación computacional de los «detectores de obstáculos» de la rana. Quizás alguien lo ha hecho ya. Pero todos sabemos que una vez que se entiende cómo *funciona efectivamente* el sistema visual de la rana, el «nivel computacional» es sencillamente irrelevante.

Para ilustrar la segunda de las cosas que deberían preocuparnos, consideremos las simulaciones de otras clases de sistemas. Por ejemplo, estoy mecanografiando estas palabras en una máquina que simula la conducta de una máquina de escribir de las antiguas.[6] Por lo que respecta a las simulaciones, el programa de procesamiento de textos simula una máquina de escribir mucho mejor de lo que simula el cerebro cualquier programa de IA. Pero ninguna persona que esté en sus cabales piensa: «al fin entendemos cómo funcionan las máquinas de escribir, son implementaciones de programas de procesamiento de textos». No es simplemente el caso que, en general, las simulaciones computacionales proporcionen explicaciones causales de los fenómenos simulados.

¿Qué pasa entonces? No suponemos en general que las simulaciones computacionales de los procesos cerebrales nos dan unas explicaciones que sustituyan a, o sean adicionales a, las explicaciones neurobiológicas de cómo funciona efectivamente el cerebro. Y, en general, no consideramos que «*X* es una simulación de *Y*» nombre una relación simétrica. Esto es: no suponemos que puesto que el ordenador simula una máquina de escribir la máquina de escribir simula consiguientemente un ordenador. No suponemos que porque un programa sobre el clima simule un huracán, entonces el programa proporciona la explicación causal de la conducta del huracán. Así pues, ¿por qué habríamos de hacer una excepción a estos principios cuando se trata de procesos cerebrales desconocidos? ¿Hay buenas razones para hacer la excepción? ¿Y qué género de explicación causal es una explicación que cita un programa formal?

Aquí, creo, está la solución a nuestro problema. Una vez que se elimina el homúnculo del sistema nos quedamos sólo con un modelo de eventos al que alguien podría adjuntar desde fuera una interpretación computacional. El único sentido en el que la especificación del modelo proporciona por sí mismo una explicación causal es este: si se sabe que existe cierto patrón en un sistema, se sabe que hay cierta causa que es

6. El ejemplo fue sugerido por John Batali.

responsable del patrón. Así se puede, por ejemplo, predecir estados posteriores a partir de estados anteriores. Además, si ya se sabe que el sistema ha sido programado por un homúnculo externo, se pueden dar explicaciones que hagan referencia a la intencionalidad del homúnculo. Se puede decir, por ejemplo, que esta máquina se comporta de la manera que se comporta porque está ejecutando un determinado programa. Esto es algo parecido a explicar que este libro empieza con unas pocas páginas sobre familias felices y no contiene ningún pasaje extenso sobre un montón de hermanos porque es *Ana Karenina* de Tolstoi y no *Los hermanos Karamazov* de Dostoievski. Pero no se puede explicar un sistema físico como una máquina de escribir o un cerebro identificando un modelo que comparte con su simulación computacional, puesto que la existencia del modelo no explica cómo funciona efectivamente el sistema *en tanto que sistema físico*. En el caso de la cognición, el modelo está en un nivel de abstracción demasiado elevado para explicar eventos mentales (y, por lo tanto, físicos) concretos tales como la ocurrencia de una percepción visual o la comprensión de una oración.

Pienso que es obvio que no podemos explicar cómo funcionan las máquinas de escribir y los huracanes señalando modelos formales que comparten con sus simulaciones computacionales. ¿Por qué esto no es obvio en el caso del cerebro?

Llegamos aquí a la segunda parte de nuestra solución del problema. Al presentar el argumento a favor del cognitivismo estábamos suponiendo tácitamente que el cerebro podría estar implementando algoritmos para la cognición en el mismo sentido en que el ordenador humano de Turing y su ordenador mecánico implementan algoritmos. Pero hemos visto que es precisamente esta suposición la que es errónea. Para ver esto, preguntémonos a nosotros mismos qué sucede cuando un sistema implementa un algoritmo. El ordenador humano recorre conscientemente los pasos del algoritmo, de modo que el proceso es a la vez causal y lógico: lógico porque el algoritmo proporciona un conjunto de reglas para derivar los símbolos de *output* a partir de los símbolos de *input*, y causal porque el agente está haciendo un esfuerzo consciente para recorrer todos los pasos. En el caso del ordenador mecánico, el funcionamiento total del sistema incluye un homúnculo externo, y con el homúnculo el sistema es a la vez causal y lógico: lógico porque el homúnculo da una interpretación a los procesos de la máquina, y causal porque el *hardware* de la máquina causa que recorra todo el proceso. Pero estas condiciones no pueden ser cumplidas por las operaciones

neurofisiológicas brutas, ciegas, no conscientes del cerebro. En el ordenador cerebral no hay implementación intencional consciente alguna del algoritmo como la hay en el ordenador humano, pero no puede haber tampoco implementación no consciente alguna como la que hay en el ordenador mecánico, puesto que esto exige un homúnculo externo que añada una interpretación computacional a los eventos físicos. Lo máximo que podemos encontrar en el cerebro es un patrón de eventos que sea formalmente similar al programa implementado en el ordenador mecánico, pero ese patrón, en cuanto tal, no tiene poderes causales a los que acudir y, por lo tanto, no explica nada.

En suma, el hecho de que la atribución de sintaxis no identifique ningún poder causal es nefasto para la pretensión de que los programas proporcionan explicaciones causales de la cognición. Para explorar las consecuencias de esto, recordemos qué aspecto tienen de hecho las explicaciones cognitivistas. Explicaciones tales como la que Chomsky ofrece de la sintaxis de los lenguajes naturales o la explicación de Marr de la visión enuncian un sistema de reglas de acuerdo con el cual un *input* simbólico se convierte en un *output* simbólico. En el caso de Chomsky, por ejemplo, un símbolo de *input* individual, *S*, se convierte en cualquiera de las potencialmente infinitas oraciones por la aplicación repetida de un conjunto de reglas sintácticas. En el caso de Marr, las representaciones de una disposición visual de dos dimensiones se convierten en «descripciones» tridimensionales del mundo de acuerdo con ciertos algoritmos. La distinción tripartita de Marr entre la tarea computacional, la solución algorítmica de la tarea y la implementación de *hardware* del algoritmo se ha convertido (como las distinciones de Newell) en el enunciado mejor conocido del modelo general de explicación.

Si se toman esas explicaciones ingenuamente, como yo hago, es mejor pensar en ellas como diciendo que es como si un hombre solo en una habitación estuviera recorriendo un conjunto de pasos consistente en seguir reglas para generar oraciones castellanas o descripciones de tres dimensiones, como podría ser el caso. Pero ahora, preguntemos qué hechos del mundo real se supone que se corresponden con esas explicaciones en tanto que aplicadas al cerebro. En el caso de Chomsky, por ejemplo, no se supone que pensamos que el agente recorre conscientemente un conjunto de aplicaciones repetidas de las reglas; ni se supone que pensamos que está pensando inconscientemente el camino que sigue a través del conjunto de reglas. Más bien, las reglas son «computacionales» y el cerebro está llevando a cabo las computaciones.

¿Pero qué significa esto? Bien, se supone que pensamos que esto es lo mismo que un ordenador comercial. Se supone que la clase de cosa que corresponde a la adscripción del mismo conjunto de reglas a un ordenador comercial corresponde a la adscripción de esas reglas al cerebro. Pero hemos visto que en el ordenador comercial la adscripción es siempre relativa al observador, la adscripción se hace de manera relativa a un homúnculo que asigna interpretaciones computacionales a los estados de *hardware*. Sin el homúnculo no hay computación, sólo hay un circuito electrónico. Así pues, ¿cómo obtenemos la computación en el cerebro sin un homúnculo? Por lo que sé ni Chomsky ni Marr han hecho frente jamás a esta cuestión ni han pensado tan siquiera que había tal cuestión. Pero sin un homúnculo, no hay poder explicativo para la postulación de estados de programa. Sólo hay un mecanismo físico, el cerebro, con diversos niveles de descripción causal físicos reales y físico-mentales.

Resumen del argumento de esta sección

La discusión de esta sección ha sido más prolija de lo que hubiera deseado, pero creo que puede resumirse rápidamente como sigue:

Objeción: es un hecho puro y simple el que las explicaciones computacionales son causales. Por ejemplo, los ordenadores pilotan aviones, y la explicación de cómo lo hacen se da en términos del programa. ¿Qué podría ser más causal que esto?

Respuesta: el sentido en el que el programa da una explicación causal es el siguiente. Hay una clase de equivalencia de los sistemas físicos tal que los modelos del sistema nos permiten la codificación de información en rasgos físicos intrínsecos del sistema, tales como los niveles de voltaje. Y esos patrones, junto con los transductores y los extremos de *input* y *output* del sistema, nos capacitan para usar cualquier miembro de esta clase de equivalencia para pilotar el avión. La generalidad de los patrones facilita las asignaciones de interpretaciones computacionales (no sorprendentemente, puesto que los patrones estaban diseñados comercialmente para este propósito), pero las interpretaciones siguen sin ser intrínsecas a los sistemas. En la medida en que la explicación hace referencia a un programa necesita un homúnculo.

Objeción: sí, pero supongamos que podemos descubrir tales patrones en el cerebro. Todo lo que la ciencia computacional cognitiva necesita es la ocurrencia de tales patrones intrínsecos.

Respuesta: desde luego que podrían descubrirse tales patrones. El cerebro tiene más patrones de los que cualquiera necesita. Pero incluso si constreñimos los patrones exigiendo las conexiones causales apropiadas y los contrafácticos consecuentes, el descubrimiento del patrón no explicaría todavía lo que estamos intentando explicar. No estamos intentando averiguar cómo un homúnculo externo podría asignar una interpretación computacional a los procesos cerebrales. Más bien, estamos intentando explicar la ocurrencia de fenómenos biológicos concretos tales como la comprensión consciente de una oración, o la experiencia visual de una escena. Esta explicación exige una comprensión de los procesos físicos sin más que producen los fenómenos.

VIII. CUARTA DIFICULTAD: EL CEREBRO NO NECESITA HACER PROCESAMIENTO DE LA INFORMACIÓN

En esta sección vuelvo finalmente a lo que pienso que es el problema central de todo esto, el problema del procesamiento de la información. Mucha gente que está dentro del paradigma científico de la «ciencia cognitiva» tendrá la sensación de que gran parte de mi discusión es simplemente irrelevante, y argumentarán en contra de ella de la manera siguiente:

> Hay una diferencia entre el cerebro y todos los demás sistemas que has estado describiendo, y esta diferencia explica por qué una simulación computacional es en el caso de los demás sistemas una mera simulación, mientras que en el caso del cerebro una simulación computacional está duplicando efectivamente y no meramente modelando las propiedades funcionales del cerebro. La razón es que el cerebro, a diferencia de esos otros sistemas, es un sistema de *procesamiento de la información*. Y este hecho sobre el cerebro es, con tus propias palabras, «intrínseco». Es pura y simplemente un hecho biológico el que el cerebro funciona para procesar la información, y puesto que podemos también procesar la misma información computacionalmente, los modelos computacionales de procesos cerebrales tienen un papel diferente de los modelos computacionales de, por ejemplo, el clima.

> Así pues, hay una pregunta bien definida para la investigación: ¿son los procesos computacionales mediante los que el cerebro procesa la información los mismos que los procesos mediante los que los ordenadores procesan la misma información?

Lo que acabo de imaginar que un oponente diría incorpora uno de los peores errores de la ciencia cognitiva. El error es suponer que los cerebros procesan información precisamente en el mismo sentido en que los ordenadores son usados para procesar información. Para ver que esto es un error sólo hay que contrastar lo que sucede en el ordenador con lo que sucede en el cerebro. En el caso del ordenador, un agente externo codifica alguna información de una forma que puede ser procesada por los circuitos del ordenador. Esto es: se proporciona una realización sintáctica de la información que el ordenador puede implementar en, por ejemplo, diferentes niveles de voltaje. El ordenador pasa entonces a través de una serie de estados eléctricos que el agente externo puede interpretar tanto sintáctica como semánticamente, aunque, desde luego, el *hardware* no tenga sintaxis o semántica intrínseca: todo está en los ojos del observador. Y la física no importa sólo en el supuesto de que se pueda lograr la implementación del algoritmo. Finalmente, se produce un *output* en forma de fenómeno físico, por ejemplo alguna muestra impresa, que un observador puede interpretar como símbolos con una sintaxis y una semántica.

Pero contrastemos esto ahora con el cerebro. En el caso del cerebro, ninguno de los procesos neurobiológicos relevantes son relativos al observador (aunque desde luego, al igual que cualquier cosa, pueden describirse desde un punto de vista relativo al observador), y la especificidad de la neurología es algo que importa absolutamente. Para dejar clara esta diferencia, examinemos un ejemplo. Supongamos que veo un coche que viene hacia mí. Un modelo computacional estándar de la visión captará información sobre la disposición visual que hay en mi retina y, eventualmente, imprimirá la oración: «Hay un coche que viene hacia mí». Pero esto no es lo que sucede en la biología efectiva. En la biología se produce una serie concreta y específica de reacciones electroquímicas en virtud del asalto de los fotones a las células fotorreceptoras de la retina, y todo este proceso da lugar eventualmente a la experiencia visual concreta. La realidad biológica no es que el sistema visual produzca todo un ramillete de palabras o símbolos; más bien se trata de un evento visual concreto específico y consciente —esta misma experiencia visual. Este

evento visual concreto es tan específico y tan concreto como un huracán o como la digestión de una comida. Podemos hacer con el ordenador un modelo de procesamiento de información de ese evento o de su producción, como podemos hacer un modelo de procesamiento de información del clima, la digestión, o de cualquier otro fenómeno, pero los fenómenos mismos no son por ello sistemas de procesamiento de la información.

Dicho brevemente, el sentido de procesamiento de la información que se usa en la ciencia cognitiva está a un nivel de abstracción demasiado elevado como para capturar la realidad biológica concreta de la intencionalidad intrínseca. La «información» del cerebro es siempre específica de una u otra modalidad. Es específica, por ejemplo, del pensamiento, o de la visión, o del oído, o del tacto. El nivel de procesamiento de la información que se describe en los modelos computacionales de cognición de la ciencia cognitiva es simplemente, por otra parte, un asunto consistente en obtener un conjunto de símbolos como *output* en respuesta a un conjunto de símbolos como *input*.

Estamos ciegos ante esta diferencia debido al hecho de que la oración «Veo un coche que viene hacia mí» puede usarse para registrar tanto la intencionalidad visual como el *output* del modelo de visión computacional. Pero esto no debería oscurecer el hecho de que la experiencia visual es un evento consciente concreto y se produce en el cerebro por procesos biológicos electroquímicos específicos. Confundir esos eventos y procesos con una manipulación de símbolos formales es confundir la realidad con el modelo. El resultado de esta parte de la discusión es que, en el sentido de «información» usado en la ciencia cognitiva, es simplemente falso decir que el cerebro es un dispositivo de procesamiento de la información.

IX. Resumen del argumento

1. De acuerdo con la definición estándar de los libros de texto, la computación se define sintácticamente en términos de manipulación de símbolos.

2. Pero sintaxis y símbolos no se definen en términos de la física. Aunque las instancias de un símbolo son siempre instancias físicas, «símbolo» y «mismo símbolo» no se definen en términos de rasgos físicos. La sintaxis no es, dicho brevemente, intrínseca a la física.

3. Esto tiene como consecuencia el que la computación no se des-

cubre en la física, se asigna a ella. Ciertos fenómenos físicos se usan, programan o interpretan sintácticamente. Sintaxis y símbolos son relativos al observador.

4. Se sigue que no se puede *descubrir* que el cerebro o cualquier otra cosa era un ordenador digital, aunque se le pueda asignar una interpretación computacional como puede hacerse con cualquier otra cosa. El asunto no es que la afirmación «El cerebro es un ordenador digital» sea simplemente falsa. Más bien, el asunto es que no alcanza el nivel de falsedad. No tiene un sentido claro. La cuestión «¿Es el cerebro un ordenador digital?» está mal definida. Si pregunta: «¿Podemos asignar una interpretación computacional al cerebro?», la respuesta es trivialmente que sí, puesto que podemos asignar una interpretación computacional a cualquier cosa. Si lo que pregunta es: «¿Son los procesos cerebrales intrínsecamente computacionales?», la respuesta es trivialmente que no, excepto, naturalmente, en el caso de los agentes conscientes que intencionalmente llevan a cabo computaciones.

5. Algunos sistemas físicos facilitan el uso computacional mucho mejor que otros. Esta es la razón por la que los construimos, los programamos y los usamos. En tales casos somos los homúnculos del sistema que estamos interpretando la física tanto en términos sintácticos como semánticos.

6. Pero las explicaciones causales que damos entonces no citan propiedades causales diferentes de las de la física de la implementación y de las de la intencionalidad del homúnculo.

7. La salida estándar, aunque tácita, de esto es cometer la falacia del homúnculo. La falacia del homúnculo es endémica en los modelos computacionales de la cognición y no puede eliminarse por argumentos estándar de descomposición recursiva. Éstos tratan de un problema diferente.

8. No podemos evitar los anteriores resultados suponiendo que el cerebro está haciendo «procesamiento de la información». El cerebro, por lo que respecta a sus operaciones intrínsecas, no hace procesamiento de la información. Es un órgano biológico específico y sus procesos neurobiológicos específicos causan formas específicas de intencionalidad. En el cerebro hay, intrínsecamente, procesos neurobiológicos y éstos causan algunas veces la conciencia. Pero este es el final de la historia. Todas las demás atribuciones mentales son o bien disposicionales, como cuando adscribimos estados inconscientes al agente, o son relativas al observador, como cuando asignamos una interpretación computacional a sus procesos cerebrales.

10. LO QUE HAY QUE ESTUDIAR

I. INTRODUCCIÓN: MENTE Y NATURALEZA

En cualquier libro sobre losofía de la mente el autor, explícita o implícitamente, tiene una visión general de la mente y de su relación con el mundo natural. El lector que ha seguido mi argumentación hasta aquí no tendrá dificultad alguna para reconocer cuál es mi visión. Veo el cerebro humano como un órgano como cualquier otro, como un sistema biológico. Su rasgo especial, por lo que respecta a la mente, el rasgo en el que difiere notablemente de otros órganos biológicos, es su capacidad de producir y mantener toda la enorme variedad de nuestra vida consciente.[1] Por conciencia no entiendo la pasividad subjetiva de la tradición cartesiana, sino todas las formas de nuestra vida consciente —desde luchar, huir, comer y fornicar a conducir coches, escribir libros o rascarnos. Todos los procesos en los que pensamos como especialmente mentales —la percepción, el aprendizaje, la inferencia, la toma de decisiones, la resolución de problemas, las emociones, etc.— están crucialmente relacionados, de una manera u otra, con la conciencia. Además, todos esos grandes rasgos que los filósofos han pensado que son especiales de la mente, dependen de forma similar de la conciencia: la subjetividad, la intencionalidad, la racionalidad, el libre albedrío (si es que hay tal cosa), y la causación mental. El olvidarse de la conciencia es lo que da cuenta, más que cualquier otra cosa, de la ausencia de frutos y la esterilidad de la psicología, la filosofía de la mente y la ciencia cognitiva.

1. El cerebro tiene también, desde luego, muchos otros rasgos que no tienen nada que ver con la conciencia. Por ejemplo, la médula regula la respiración incluso cuando el sistema está totalmente inconsciente.

El estudio de la mente es el estudio de la conciencia en el mismo sentido en que la biología es el estudio de la vida. La biología no necesita, desde luego, estar pensando constantemente sobre la vida y, de hecho, la mayor parte de los escritos sobre biología no necesitan tan siquiera usar el concepto de vida. Sin embargo, nadie que esté en sus cabales niega que los fenómenos que estudian en biología son formas de vida. Ahora bien, de forma similar, el estudio de la mente es el estudio de la conciencia, incluso aunque uno pueda no hacer explícitamente mención alguna de la conciencia cuando se está haciendo un estudio de la inferencia, la percepción, la toma de decisiones, la resolución de problemas, los actos de habla, etc.

Nadie puede, o debe, intentar predecir o legislar el futuro de la investigación en filosofía, ciencia u otras disciplinas. Nos sorprenderá el hallazgo de nuevos conocimientos y una de las sorpresas que deberíamos esperar es que los avances en el conocimiento no nos darán sólo nuevas explicaciones, sino también nuevas *formas* de explicación. En el pasado, por ejemplo, la revolución darwiniana produjo un nuevo tipo de explicación, y creo que no hemos entendido completamente su importancia para nuestra situación presente.

En este capítulo final quiero explorar algunas de las consecuencias de la posición filosófica general que he estado defendiendo respecto de la filosofía de la mente. Comienzo con una discusión del principio de conexión y sus implicaciones.

II. LA INVERSIÓN DE LA EXPLICACIÓN

Creo que el principio de conexión tiene algunas consecuencias totalmente sorprendentes. Argumentaré que muchas de nuestras explicaciones en ciencia cognitiva carecen de la fuerza explicativa que pensábamos que tenían. Para rescatar lo que se puede salvar de ellas, tendremos que realizar una inversión en su estructura lógica análoga a la inversión que los modelos darwinianos de la explicación biológica impusieron a la vieja biología teleológica que precedió a Darwin.

En nuestros cráneos lo único que tenemos es el cerebro con toda su intrincada estructura, y la conciencia con todo su color y variedad. El cerebro produce los estados conscientes que ocurren en nosotros ahora mismo, y tiene la capacidad de producir muchos otros que no están ocurriendo ahora mismo. Pero esto es lo que hay. Por lo que respecta a

la mente, este es el final de la historia. Hay procesos neurofisiológicos brutos, ciegos, y hay conciencia, pero no hay nada más. Si estamos buscando fenómenos que sean intrínsecamente intencionales pero inaccesibles en principio a la conciencia, no hay nada de eso: no hay seguimiento de reglas, no hay procesamiento de la información mental, no hay bosquejos primigenios, no hay imágenes de dos dimensiones y media, no hay descripciones tridimensionales, no hay lenguaje del pensamiento, y no hay gramática universal. En lo que sigue argumentaré que toda la historia cognitivista que postula todos esos fenómenos mentales inaccesibles está basada en una concepción predarwiniana de la función del cerebro.

Consideremos el caso de las plantas y las consecuencias de la revolución darwiniana sobre el aparato explicativo que usamos para dar cuenta de su conducta. Antes de Darwin, era bastante común antropomorfizar la conducta de las plantas y decir cosas tales como que las plantas vuelven sus hojas hacia el sol para ayudarse en su supervivencia. La planta «quiere» sobrevivir y florecer, y «para lograrlo» sigue al sol. De acuerdo con esta concepción predarwiniana se suponía que había un nivel de intencionalidad en la conducta de la planta. Este nivel de supuesta intencionalidad ha sido reemplazado ahora por otros dos niveles de explicación, un nivel de «*hardware*» y un nivel «funcional». En el nivel de *hardware* hemos descubierto que los movimientos efectivos de las hojas que la planta realiza al seguir al sol están causados por una hormona específica, la auxina. Las secreciones variables de auxina dan cuenta de la conducta de la planta sin ninguna hipótesis o propósito, teleología o intencionalidad adicionales. Obsérvese, además, que esta conducta juega un papel crucial en la supervivencia de la planta, de modo que en el nivel funcional podemos decir cosas tales como que la conducta de buscar la luz que la planta manifiesta funciona para ayudar a la planta a sobrevivir y a reproducirse.

La explicación intencional original de la conducta de la planta resultó ser falsa, pero no sólo era falsa. Si nos desembarazamos de la intencionalidad e invertimos el orden de la explicación, la afirmación de intencionalidad aparece intentando decir algo verdadero. Para que lo que sucede resulte completamente claro, quiero mostrar cómo al reemplazar la explicación intencional original por una combinación de la explicación del *hardware* mecánico y una explicación funcional, estamos invirtiendo la estructura explicativa de la explicación intencional original.

a. La explicación intencional original:
Puesto que quiere sobrevivir, la planta vuelve sus hojas hacia el sol.
o:
Para sobrevivir, la planta vuelve sus hojas hacia el sol.
b. La explicación del *hardware* mecánico:
Las secreciones variables de auxina causan que las plantas vuelvan sus hojas hacia el sol.
c. La explicación funcional:
Las plantas que vuelven sus hojas hacia el sol *es más probable que sobrevivan que las plantas que no lo hacen.*

En (*a*) la forma de explicación es teleológica. La *representación* de la meta, esto es: la supervivencia, funciona como la *causa* de la conducta, a saber: volverse hacia el sol. Pero en (*c*) la teleología se elimina y la conducta que ahora, de acuerdo con (*b*), tiene una explicación mecánica, causa el hecho bruto de la supervivencia, que ya no es una meta, sino sólo un efecto que, simplemente, sucede.

La moraleja que más adelante extraeré de toda esta discusión puede enunciarse, al menos en una forma preliminar, de la manera siguiente: *por lo que respecta a los procesos no conscientes, estamos todavía antropomorfizando el cerebro de la misma manera en que estábamos antropomorfizando las plantas antes de la revolución darwiniana.* Es fácil ver por qué podemos cometer el error de antropomorfizar el cerebro —despues de todo el cerebro es el hogar de *anthropos*. Sin embargo, adscribir una amplia disposición de fenómenos intencionales a un sistema en el que se violan las condiciones de esa adscripción es simplemente erróneo. Lo mismo que la planta no tiene estados intencionales puesto que no reúne las condiciones para tener estados intencionales, del mismo modo aquellos procesos cerebrales que no son en principio accesibles a la conciencia no tienen intencionalidad, porque no reúnen condiciones para tener intencionalidad. Cuando adscribimos intencionalidad a procesos del cerebro que son inaccesibles en principio a la conciencia, lo que decimos es o bien metafórico —como en las adscripciones metafóricas de estados mentales a las plantas— o falso. Las adscripciones de intencionalidad a las plantas serían falsas si las tomáramos literalmente. Pero téngase en cuenta que no son *pura y simplemente* falsas; están intentando decir algo verdadero, y para llegar a lo que hay en ellas de verdadero tenemos que invertir muchas de las explicaciones de la ciencia cognitiva como hicimos en biología de las plantas.

Para elaborar estas tesis con detalle, tenemos que considerar algunos casos específicos. Empezaré con teorías de la percepción y luego continuaré con teorías del lenguaje para mostrar qué aspecto podría tener una ciencia cognitiva que respetase los hechos del cerebro y los hechos de la conciencia.

Irving Rock concluye su excelente libro sobre la percepción (Rock, 1984) con las siguientes observaciones: «Aunque la percepción es autónoma respecto a facultades mentales superiores tales como las que se exhiben en el pensamiento consciente y en el uso de conocimiento consciente, argumentaré a pesar de todo que es inteligente. Cuando llamo "inteligente" a la percepción lo que quiero decir es que se basa en procesos mentales similares al pensamiento tales como la descripción, la inferencia, la resolución de problemas, aunque esos procesos sean veloces como el rayo, inconscientes y no verbales ... "Inferencia" implica que ciertas propiedades perceptivas se computan a partir de información sensorial dada usando reglas conocidas inconscientemente. Por ejemplo: el tamaño percibido se infiere a partir del ángulo visual del objeto, su distancia percibida y la ley de la óptica geométrica que relaciona el ángulo visual con la distancia del objeto» (p. 234).

Apliquemos ahora esta tesis, a modo de ejemplo, a la explicación de la ilusión de Ponzo.

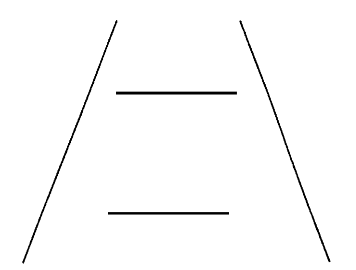


FIGURA 10.1. *La ilusión de Ponzo.*

Aunque las dos líneas paralelas tienen la misma longitud, la línea superior parece más larga. ¿Por qué? De acuerdo con la explicación estándar, el agente está siguiendo inconscientemente dos reglas y está haciendo dos inferencias inconscientes. La primera regla es que las líneas convergentes de abajo hacia arriba que están en el campo visual impli-

can una mayor distancia en la dirección de la convergencia, y la segunda es que aquellos objetos que ocupan porciones de la imagen retiniana varían en su tamaño percibido dependiendo de la distancia percibida desde el observador (ley de Emmert). De acuerdo con esta explicación, el agente infiere inconscientemente que la línea paralela superior está más lejos a causa de su posición en relación con las líneas convergentes y, en segundo lugar, infiere que la línea superior es más larga porque está más lejos. Así pues, hay dos reglas y dos inferencias inconscientes, y ninguna de estas operaciones es accesible a la conciencia, ni siquiera en principio. Debe señalarse que esta explicación es controvertida y que han surgido muchísimas objeciones en su contra (véase Rock, 1984, pp. 156 y ss.). Pero el asunto importante es aquí que no se desafía la *forma* de la explicación que es lo que yo estoy desafiando en este momento. Estoy interesado en este tipo de explicación y no sólo en los detalles de este ejemplo.

No hay modo alguno en el que este tipo de explicación pueda hacerse consistente con el principio de conexión. Puede verse esto si nos preguntamos: «¿Qué hechos del cerebro se supone que corresponden a la adscripción de todos estos procesos mentales inconscientes?». Sabemos que hay experiencias visuales conscientes, y sabemos que están causadas por procesos cerebrales, pero ¿dónde se supone que está en este caso el nivel mental adicional? De hecho, este ejemplo es muy difícil de interpretar literalmente sin hacer uso de un homúnculo: estamos postulando operaciones lógicas realizadas sobre imágenes retinianas, pero ¿quién se supone que está realizando esas operaciones? Un examen detallado revela que esta explicación es, en su misma forma, la antropomorfización de procesos no conscientes del cerebro en el mismo sentido en que las explicaciones predarwinianas de la conducta de las plantas habían antropomorfizado las operaciones no conscientes de la planta.

El problema no es, como se afirma algunas veces, que carecemos de evidencia empírica suficiente para la postulación de procesos mentales que son, en principio, inaccesibles a la conciencia; más bien no está claro qué se supone lo que significa la postulación. No podemos hacerla coherente con que sabemos acerca de los estados mentales y lo que sabemos sobre las operaciones del cerebro. Pensamos, en nuestra ignorancia patética sobre el funcionamiento del cerebro, que algún día una ciencia del cerebro avanzada les buscará un lugar a todos esos procesos inteligentes e inconscientes. Pero uno sólo tiene que imaginar los

detalles de una ciencia perfecta del cerebro para ver que, incluso si tuviéramos tal ciencia, no habría lugar en ella para la postulación de tales procesos. Una ciencia perfecta del cerebro estaría enunciada en vocabulario neurofisiológico (esto es: de *hardware*). Habría, por lo que respecta al *hardware*, distintos niveles de descripción y, lo mismo que en el caso de las plantas, habría también niveles funcionales de descripción. Estos niveles funcionales identificarían aquellos rasgos del *hardware* que encontramos interesantes del mismo modo que nuestras descripciones funcionales de la planta identifican aquellas operaciones de *hardware* en las que tenemos algún interés. Pero lo mismo que la planta no sabe nada sobre la supervivencia, tampoco las operaciones no conscientes del cerebro saben nada de inferencias, seguir reglas o de juicios sobre tamaño y distancia. Atribuimos esas funciones al *hardware* de manera relativa a nuestros intereses, pero no hay hecho mental adicional de ninguna clase que esté involucrado en las atribuciones funcionales.

La diferencia crucial entre el cerebro, de un lado, y la planta, de otro, es esta: el cerebro tiene intrínsecamente un nivel mental de descripción puesto que en un momento dado está causando eventos conscientes efectivos y es capaz de causar eventos conscientes posteriores. Puesto que el cerebro tiene estados mentales tanto conscientes como inconscientes, estamos también tentados a suponer que en el cerebro hay estados mentales que son inaccesibles a la conciencia. Pero esta tesis es inconsistente con el principio de conexión, y aquí necesitamos hacer la misma inversión de la explicación que hemos hecho en las explicaciones de la conducta de la planta. En lugar de decir: «Vemos más larga la línea superior porque estamos siguiendo inconscientemente dos reglas y haciendo dos inferencias», deberíamos decir: «Vemos conscientemente la línea superior como si estuviese más lejos y fuese más larga». Punto. Fin de la historia intencional.

Como sucede con la planta, hay aquí una historia funcional y una historia (bastante desconocida) de *hardware* mecánico. El cerebro funciona de tal manera que las líneas convergentes anteriores parecen alejarse de nosotros en dirección a la convergencia, y los objetos que producen el mismo tamaño de imagen retiniana parecen variar de tamaño si se percibe que están a distancias diferentes de nosotros. *Pero en este nivel funcional no hay contenido mental de ningún género.* En tales casos el sistema funciona para causar ciertas clases de intencionalidad consciente, pero la causación no es ella misma intencional. Y lo impor

tante, para repetirlo otra vez, no es que la adscripción de intencionalidad inconsciente profunda esté insuficientemente apoyada por la evidencia empírica, sino que no se puede hacer coherente con lo que ya sabemos que es el caso.

«Bien —podría decirse—, la distinción no importa realmente demasiado a la ciencia cognitiva. Continuamos diciendo lo que siempre hemos dicho y continuamos haciendo lo que siempre hemos hecho: simplemente sustituimos en estos casos la palabra 'mental' por la palabra 'funcional'. Se trata de una sustitución que, en cualquier caso, la mayoría de nosotros hemos estado haciendo inconscientemente, puesto que muchos de nosotros tendemos a usar estas palabras de modo que sean intercambiables.»

Pienso que la afirmación que estoy haciendo tiene importantes aplicaciones para la investigación en ciencia cognitiva, puesto que al invertir el orden de la explicación obtenemos una explicación diferente de las relaciones causa-y-efecto y, al hacerlo así, alteramos radicalmente la estructura de la explicación psicológica. En lo que sigue tengo dos propósitos: quiero desarrollar la afirmación original de que la ciencia cognitiva exige una inversión de la explicación comparable a la inversión lograda por la biología evolucionista, y quiero mostrar algunas de las consecuencias que tendría esta inversión para la conducción de nuestra investigación.

Creo que el error persiste en gran medida porque, en el caso del cerebro, carecemos de explicaciones de *hardware* del tipo de la auxina. Quiero explicar la inversión en un caso en el que tenemos algo parecido a una explicación de *hardware*. Cualquiera que haya visto una cinta de vídeo doméstico filmada desde un coche en movimiento se sorprende de hasta qué punto el mundo se mueve mucho más en la cinta de lo que lo hace en la vida real. ¿Por qué? Imaginémonos que vamos en coche por una carretera llena de baches. Uno mantiene fijos sus ojos de manera consciente en la carretera y en el tráfico restante aun cuando el coche y lo que hay dentro de él, incluyendo el cuerpo de uno mismo, vayan dando botes. Además de los esfuerzos conscientes de mantener los ojos en la carretera, está sucediendo algo más de modo inconsciente: los globos oculares se están moviendo constantemente dentro de sus cuencas de manera que ayuden a mantener la visión concentrada en la carretera. Se puede intentar hacer el experimento del modo siguiente: fija tu mirada en la página que tienes frente a ti y mueve tu cabeza a derecha e izquierda y de arriba abajo.

En el caso del coche resulta tentador pensar que estamos siguiendo una regla inconsciente. Una primera aproximación a esta regla sería: mueve los globos oculares en las cuencas de manera relativa al resto de la cabeza de modo que la mirada se mantenga fija sobre el objeto en el que intentas mantenerla. Obsérvese que las predicciones de esta regla no son triviales. Otro modo de hacerlo habría sido mantener los ojos fijos en sus cuencas y mover la cabeza; de hecho algunos pájaros mantienen de esta manera su estabilidad retiniana. (Si una lechuza pudiera conducir un coche, esta sería la manera que tendría de hacerlo, puesto que sus globos oculares están fijos.) Así pues, tenemos dos niveles de intencionalidad:

Una intención consciente: mantén tu atención visual en la carretera.
Una regla inconsciente profunda: haz movimientos de los globos oculares en relación con sus cuencas que sean iguales y opuestos a los movimientos de la cabeza para mantener estable la imagen retiniana.

En este caso, el resultado es consciente, aunque los medios para lograrlo sean inconscientes. Pero el aspecto inconsciente tiene todas las señales de la conducta inteligente. Es complejo, flexible, se dirige a una meta, incluye procesamiento de la información y tiene una capacidad generativa potencialmente infinita. Esto es: el sistema recibe información sobre movimientos corporales e imprime instrucciones sobre los movimientos de los globos oculares, sin ningún límite sobre las combinaciones posibles de los movimientos de los globos oculares que el sistema puede generar. Además, el sistema puede aprender porque la regla puede modificarse sistemáticamente poniéndole al agente lentes de aumento o de disminución. Y sin mucha dificultad, uno podría contar cualquier historia de ciencia cognitiva estándar sobre la conducta inconsciente: una historia sobre procesamiento de la información, el lenguaje del pensamiento, y los programas de ordenador, por sólo mencionar algunos ejemplos obvios. Dejo al lector esto a modo de ejercicio para que construya la historia de acuerdo con su paradigma favorito de ciencia cognitiva.

El problema es, sin embargo, que todas estas historias son falsas. Lo que sucede efectivamente es que los movimientos de fluidos en los canales semicirculares del oído interno provocan una secuencia de disparos de neuronas que entran en el cerebro a través del octavo nervio craneal. Estas señales siguen dos caminos paralelos, uno de los cuales

puede «aprender» mientras que el otro no. Los caminos están en la base del cerebro y en el cerebelo y transforman las señales iniciales de *input* para proporcionar «órdenes» motoras de *output*, a través de las neuronas motoras que conectan los músculos del ojo y que causan los movimientos del globo ocular. El sistema total contiene un mecanismo de retroactivación para la corrección de errores. Se le denomina reflejo ocular vestibular (ROV).[2] El mecanismo efectivo de *hardware* del ROV no tiene más intencionalidad o inteligencia que el movimiento de las hojas de la planta debido a las secreciones de auxina. La apariencia de que ahí se está siguiendo una regla inconscientemente, se está procesando información inconscientemente, etc., es una ilusión óptica. Todas las adscripciones intencionales son *como-si*. Así pues, he aquí cómo se produce la inversión de la explicación. En vez de decir:

Intencional: para mantener mi imagen retiniana estable y, por consiguiente, mejorar mi visión mientras mi cabeza se está moviendo, sigo la regla inconsciente profunda del movimiento del globo ocular.

Diríamos:

Hardware: cuando miro al objeto mientras que mi cabeza se está moviendo, el mecanismo de *hardware* del ROV mueve mis globos oculares.

Funcional: el movimiento ROV mantiene estable la imagen retiniana y esto mejora mi visión.

¿Por qué es tan importante este cambio? En las explicaciones científicas, intentamos de modo característico decir exactamente qué causa qué. En los paradigmas tradicionales de la ciencia cognitiva se supone que hay una causa mental inconsciente profunda que se supone que produce un efecto deseado como, por ejemplo, juicios perceptivos u oraciones gramaticales. Pero la inversión elimina completamente esta causa mental. No hay nada ya excepto un mecanismo físico bruto que produce un efecto físico bruto. Esos efectos y mecanismos son describibles en diferentes niveles, ninguno de los cuales es hasta ahora mental. El aparato del ROV funciona para mejorar la eficiencia visual, pero la única intencionalidad es la percepción consciente del objeto. El res-

2. Lisberger (1988); Lisberger y Pavelko (1988).

to del trabajo lo realiza en su totalidad el mecanismo físico bruto del ROV. Así pues, la inversión altera radicalmente la ontología de la explicación en ciencia cognitiva *eliminando un nivel completo de causas psicológicas inconscientes*. El elemento normativo que se suponía que estaba *dentro del sistema* en virtud de su contenido psicológico vuelve a entrar ahora cuando *un agente consciente fuera del mecanismo hace juicios sobre su funcionamiento*. Para clarificar este último punto he de decir más cosas sobre las explicaciones funcionales.

III. La lógica de las explicaciones funcionales

Podría parecer que estoy proponiendo que hay aquí, sin problematizarlo, tres niveles diferentes de explicación —*hardware*, funcional e intencional— y que donde entran en juego procesos inconscientes profundos deberíamos sustituir simplemente las explicaciones intencionales por las de *hardware* y las funcionales. Pero de hecho la situación es un poco más complicada que esto. Allí donde entran en juego las explicaciones funcionales, la metáfora de los niveles es de alguna manera desorientadora, porque sugiere que hay un nivel funcional separado que es diferente de los niveles causales. Esto no es verdad. El denominado «nivel funcional» no es un nivel separado en absoluto, sino simplemente uno de los niveles causales *descrito en términos de nuestros intereses*. En el caso de artefactos e individuos biológicos, nuestros intereses son tan obvios que pueden parecer inevitables, y el nivel funcional puede parecer intrínseco al sistema. Después de todo, ¿quién podría negar, por ejemplo, que el corazón *funciona* para bombear sangre? Pero recuérdese que cuando decimos que el corazón funciona para bombear sangre, los únicos hechos en cuestión son que el corazón, de hecho, bombea sangre; el hecho es importante para nosotros y está relacionado causalmente con todo un conjunto de otros hechos que también son importantes, como el hecho de que el bombeo de la sangre es necesario para estar vivo. Si la única cosa que nos interesase sobre el corazón fuese que hacía un ruido a base de latidos o que ejercía atracción gravitatoria sobre la Luna, entonces tendríamos una concepción completamente diferente de su «funcionamiento» y correspondientemente de, por ejemplo, las enfermedades cardiacas. Para señalar esto de manera que todo el mundo lo entienda: el corazón no tiene función alguna además de sus diversas relaciones causales. Cuando hablamos

de sus funciones, estamos hablando sobre aquellas de sus relaciones causales a las que concedemos alguna importancia *normativa.* Así pues, la eliminación del nivel inconsciente profundo señala dos cambios importantísimos: se desembaraza de todo el nivel de la causación psicológica y traslada el componente normativo del mecanismo al ojo del poseedor del mecanismo. Obsérvese, por ejemplo, el vocabulario normativo que usa Lisberger para caracterizar la función del ROV. «La función del ROV consiste en estabilizar las imágenes retinianas para generar movimientos suaves del ojo que sean iguales y opuestos a cada movimiento de la cabeza.» Además, «es importante un ROV exacto puesto que necesitamos imágenes retinianas estables para tener una buena visión» (Lisberger 1988, pp. 728-729).

El nivel intencional difiere, por otro lado, de los niveles funcionales no intencionales. Aunque ambos son causales, los rasgos causales de la intencionalidad intrínseca combinan lo causal con lo normativo. Los fenómenos intencionales tales como seguir reglas y actuar de acuerdo con deseos y creencias son fenómenos genuinamente causales; pero en tanto que fenómenos intencionales están esencialmente relacionados con fenómenos normativos tales como verdad y falsedad, éxito y fracaso, consistencia e inconsistencia, racionalidad, ilusión y, generalmente, condiciones de satisfacción.[3] Dicho brevemente: los hechos efectivos de la intencionalidad contienen elementos normativos, pero, por lo que respecta a las explicaciones funcionales, los únicos *hechos* son hechos «brutos», hechos físicos ciegos, y las únicas normas están en nosotros y existen sólo desde nuestro punto de vista.

El abandono de la creencia en una extensa clase de fenómenos mentales que son, en principio, inaccesibles a la conciencia, debería tener como resultado tratar al cerebro como un órgano igual que cualquier otro. Al igual que cualquier otro órgano, el cerebro tiene un nivel funcional —de hecho, muchos niveles funcionales— de descripción, y al igual que cualquier otro órgano *puede describirse como si* estuviese llevando a cabo «procesamiento de la información» e implementando un número cualquiera de programas de ordenador. Pero el rasgo verdaderamente especial del cerebro, el rasgo que lo convierte en el órgano de lo mental, es su capacidad para causar y mantener pensamientos conscientes, experiencias, acciones, recuerdos, etc.

La noción de *proceso* mental inconsciente y la noción correlacio-

3. Véase Searle (1983), especialmente el capítulo 5, para una discusión más extensa.

nada de los principios de los procesos mentales inconscientes son también una fuente de confusión. Si pensamos que un proceso consciente es «puramente» mental, podríamos pensar en algo parecido a tararearse a uno mismo en silencio una melodía dentro de la propia cabeza. Aquí hay claramente un proceso que tiene un contenido mental. Pero hay también un sentido de «proceso mental» que no significa «proceso con contenido mental», sino más bien «proceso mediante el que se relacionan los fenómenos mentales». En este segundo sentido, los procesos pueden tener o no un contenido mental. Por ejemplo, en la vieja psicología asociacionista se suponía que había un proceso por medio del cual la percepción de *A* me recuerda *B*, y este proceso funciona de acuerdo con el principio de semejanza. Si veo *A*, y *A* se parece a *B*, entonces tendré una tendencia a formarme una imagen de *B*. En este caso, el proceso mediante el que paso de la percepción de *A* a la imagen de *B* no incluye necesariamente ningún contenido mental adicional en absoluto. Se supone que hay un principio de acuerdo con el cual funciona el proceso, a saber: la semejanza; pero la existencia del proceso que funciona de acuerdo con el principio no implica que tenga que haber contenido mental adicional alguno que no sea la percepción de *A* y el pensamiento de *B*, o el pensamiento de que *B* se parece a *A*. En particular, no implica que cuando se ve *A* y uno se acuerda de *B*, se sigue una regla que exige que si veo *A* y si *A* se parece a *B*, entonces debería pensar en *B*. Dicho brevemente: *el proceso mediante el que se relacionan los contenidos mentales no necesita tener contenido mental alguno que sea adicional al de los* relata; ahora bien, nuestro modo de hablar teórico y nuestros pensamientos sobre ese principio tendrán, desde luego, un contenido que se refiere al principio. Esta distinción va a mostrarse como importante, puesto que muchas de las discusiones en ciencia cognitiva se mueven desde la afirmación de que hay procesos que son «mentales» en el sentido de que causan fenómenos conscientes (los procesos del cerebro que producen, por ejemplo, experiencias visuales) hasta la de que esos procesos son procesos mentales en el sentido de que tienen contenido mental, información, inferencia, etc. Los procesos no conscientes del cerebro que causan las experiencias visuales son ciertamente mentales en un sentido, pero no tienen contenido mental en absoluto y entonces, en este sentido, no son procesos mentales.

Para clarificar esta distinción, distingamos entre aquellos procesos, tales como el seguir reglas, que tienen un contenido mental que funcio-

na causalmente en la producción de la conducta, y aquellos procesos que no tienen contenido mental, pero que asocian contenidos mentales con estímulos de *input*, conducta de *output*, y otros contenidos mentales. Llamaré al último caso «patrones de asociación». Si, por ejemplo, siempre que como mucha pizza tengo dolor de estómago, estoy definitivamente ante un patrón de asociación, pero no estoy siguiendo reglas. No sigo una regla: cuando comes mucha pizza, tienes dolor de estómago; simplemente así es como sucede.

IV. Algunas consecuencias: gramática universal, patrones de asociación y conexionismo

Es característico de las explicaciones intencionales de la conducta humana y animal el que los *patrones* de la conducta se expliquen por el hecho de que el agente tiene una representación de ese mismo modelo en el aparato intencional, y que esa representación funciona causalmente en la producción del patrón de conducta. Así decimos que los británicos conducen por la izquierda porque siguen la regla siguiente: conduce por la izquierda; y que no conducen por la derecha porque siguen esa misma regla. El contenido intencional funciona causalmente al producir la conducta que representa. Hay dos puntualizaciones que hacer de manera inmediata. En primer lugar, el contenido intencional de la regla no produce en absoluto la conducta por sí mismo. Nadie, por ejemplo, coge el coche justamente para seguir la regla, y nadie habla justamente por mor de seguir las reglas del castellano. Y en segundo lugar, las reglas, principios, etc., pueden ser inconscientes y, para todos los propósitos prácticos, no están disponibles a menudo para la conciencia, aun cuando, como hemos visto, si hay realmente tales reglas, tienen que ser, al menos en principio, accesibles a la conciencia.

Una estrategia típica en ciencia cognitiva ha sido intentar descubrir patrones complejos tales como aquellos que se encuentran en la percepción o en el lenguaje y postular a continuación combinaciones de representaciones mentales que explicaran el modelo de modo apropiado. Donde no hay representación consciente o superficialmente inconsciente, postulamos una representación mental inconsciente profunda. Epistémicamente, la existencia de los modelos se toma como evidencia a favor de la existencia de las representaciones. Causalmente, se supone que la existencia de las representaciones explica la existencia de los

patrones. Pero tanto la afirmación epistémica como la causal presuponen que la ontología de las reglas inconscientes profundas está perfectamente en orden tal como está. He intentado desafiar la ontología de las reglas inconscientes profundas, y si este desafío tiene éxito, las pretensiones epistémicas y causales se derrumbarán a la vez. Epistémicamente, tanto la planta como el ROV exhiben patrones sistemáticos, pero esto no proporciona evidencia alguna a favor de la existencia de reglas inconscientes profundas —un punto que es obvio en el caso de la planta, menos obvio pero aún verdadero en el caso de la visión. Causalmente, el patrón de conducta juega un papel funcional en la conducta total del sistema, pero la representación del patrón en nuestra teoría no identifica una representación inconsciente profunda que juegue algún papel causal en la producción del patrón de conducta, porque no hay tal representación inconsciente profunda. De nuevo, este es un punto obvio en el caso de la planta, menos obvio pero todavía verdadero en el caso de la visión.

Ahora bien, con este aparato a mano, volvamos a la discusión sobre el estatus de las pretendidas reglas de gramática universal. Concentro mi atención en la gramática universal porque las gramáticas de los lenguajes particulares como el francés o el castellano contienen, además de cualquier otra cosa que contengan, un extenso número de reglas que son accesibles a la conciencia. El argumento tradicional a favor de la existencia de la gramática universal puede enunciarse de manera muy simple del modo siguiente: el hecho de que todos los niños normales puedan adquirir fácilmente el lenguaje de la comunidad en la que crecen sin una instrucción especial y sobre la base de estímulos muy imperfectos y degenerados, y además esos niños puedan aprender ciertas clases de lenguajes, tales como las ejemplificadas por los lenguajes naturales humanos, pero no puedan aprender todas las clases de otros sistemas de lenguajes lógicamente posibles, produce una evidencia abrumadora a favor de que todo niño normal, de alguna manera que nos resulta desconocida, contiene en su cerebro un dispositivo especial de adquisición del lenguaje (DEAL), y *este dispositivo de adquisición del lenguaje consiste, al menos en parte, en un conjunto de reglas inconscientes profundas.*

Con la excepción de la última oración que va en cursiva, estoy de acuerdo con el argumento anterior a favor de un «dispositivo de adquisición del lenguaje». El único problema es la postulación de reglas inconscientes profundas. Este papel es inconsistente con el principio de

conexión. No es sorprendente que haya habido bastante controversia sobre las clases de evidencia que uno debería tener a favor de la existencia de estas reglas. Estas discusiones son siempre inconclusivas, porque la hipótesis es vacía.

Hace años planteé dudas epistémicas sobre la confianza de Chomsky en la atribución de reglas inconscientes profundas y sugerí que cualquier atribución exigiría evidencia de que el contenido específico de la regla, el contorno de aspecto específico, estaba jugando un papel causal en la producción de la conducta en cuestión (Searle, 1976). Afirmé que predecir los patrones correctos no sería suficiente para justificar la pretensión de que estamos siguiendo reglas inconscientes profundas; además necesitaríamos la evidencia de que la regla era «causalmente eficaz» en la producción del patrón. Con ciertas puntualizaciones, Chomsky acepta las exigencias. Puesto que estamos de acuerdo en esas exigencias, merecería la pena enunciarlas:

1. El uso de la palabra «regla» no es importante. El fenómeno en cuestión podría ser, en principio, o un parámetro, o una constricción, o... y así sucesivamente. La cuestión importante es, sin embargo, que esto sucede al nivel de la intencionalidad. Tanto para Chomsky como para mí, no es meramente un asunto de que el sistema se comporte *como si* estuviese siguiendo una regla. Tiene que haber una diferencia entre el papel de las reglas en la facultad del lenguaje y, por ejemplo, el papel de las «reglas» en la conducta de las plantas y de los planetas.

2. La «conducta» no está aquí en disputa. La comprensión de oraciones, las intuiciones de gramaticalidad, y las manifestaciones de competencia lingüística son, en general, aquello a lo que nos estamos refiriendo mediante la abreviatura «conducta». No hay ningún conductismo implícito en el uso de este término y tampoco hay ninguna confusión entre competencia y actuación.

3. Ninguno de los dos supone que toda la conducta (en el sentido relevante) esté causada por las reglas (en el sentido relevante). El asunto importante es, sin embargo, que en la mejor explicación causal de los fenómenos las reglas «entran dentro de» (expresión de Chomsky) la teoría que da la explicación.

Ahora bien, ¿cuál fue exactamente la respuesta de Chomsky a esta objeción teniendo presentes estas constricciones?

> Supongamos que nuestro modo de explicación y descripción que tiene más éxito atribuye a Jones un estado inicial y alcanzado que inclu-

> ye ciertas reglas (principios con parámetros fijos o reglas de otras clases) y explica la conducta de Jones en estos términos; esto es: las reglas forman una parte central de la mejor explicación de su uso y comprensión del lenguaje y se invocan de manera directa y crucial al proporcionar explicaciones en la mejor teoría que podemos idear ... No puedo ver que al atribuir eficacia causal a las reglas esté involucrado algo que vaya más allá de la afirmación de que esas reglas son elementos constitutivos de los estados postulados en la teoría explicativa de la conducta y de que entran en nuestra mejor explicación de esta conducta (Chomsky 1986, pp. 252-253).

En conexión con esto mismo, Chomsky cita también a Demopoulos y Matthews (1983):

> Como Demopoulos y Matthews (1983) han observado, «la aparente indispensabilidad teórica de las apelaciones a estados internos gramaticalmente caracterizados en la explicación de conducta lingüística es seguramente la mejor clase de razón para atribuir a esos estados [y, podemos añadir, a sus elementos constituyentes relevantes] un papel causal en la producción de la conducta» (Chomsky, 1986, p. 257).

Así pues, la idea es la siguiente: la afirmación de que las reglas son causalmente eficaces se justifica por el hecho de que las reglas son elementos constituyentes de los estados postulados por la mejor teoría causal de la conducta. La objeción que quiero hacer a esta explicación debería ser obvia ya: al enunciar que la «mejor teoría» requiere la postulación de reglas inconscientes profundas de gramática universal, estos tres autores están presuponiendo que la postulación de tales reglas es, para empezar, perfectamente legítima. Pero una vez que arrojamos dudas sobre la legitimidad de esta suposición, parece entonces que la «mejor teoría» podría también tratar la evidencia como patrones de asociación que no son producidos por las representaciones mentales que, de alguna manera, reflejan esos patrones, sino que son producidos por estructuras neurofisiológicas que no necesitan tener semejanza alguna con los patrones. El *hardware* produce patrones de asociación, en el sentido definido antes, pero los patrones de asociación no juegan ningún papel causal en la producción de los patrones de conducta —son sólo esos patrones de conducta.

Específicamente, se da cuenta de manera mucho más simple de la evidencia a favor de la gramática universal por medio de la siguiente

hipótesis: de hecho, hay un dispositivo innato de adquisición del lenguaje en los cerebros humanos y DEAL constriñe la forma de los lenguajes que los seres humanos pueden aprender. Hay, pues, un nivel de explicación de *hardware* en términos de la estructura del dispositivo, y hay un nivel funcional de explicación que describe las clases de lenguaje que pueden ser adquiridas por un niño en aplicación de este mecanismo. No se añade ningún poder predictivo o explicativo nuevo diciendo que hay además un nivel de reglas de gramática universal inconscientes profundas y, de hecho, he intentado sugerir que esta postulación es, en cualquier caso, incoherente. Supongamos, por ejemplo, que los niños pueden sólo aprender lenguajes que contienen alguna propiedad formal específica *F*. Ahora bien, es evidente que DEAL hace posible el aprender lenguajes *F* y no hace posible el aprender lenguajes no *F*. Pero esto es todo. No hay ninguna evidencia de que el niño tenga una regla inconsciente profunda de este tipo: «Aprende los lenguajes *F* y no aprendas los lenguajes no *F*». Y en cualquier caso no se le ha dado ningún sentido a esta suposición.

La situación es exactamente análoga a la siguiente: los seres humanos son capaces de percibir colores sólo dentro de un cierto rango del espectro. Sin un entrenamiento formal, pueden ver, por ejemplo, azul y rojo, pero no pueden ver infrarrojo o ultravioleta. Esto constituye una evidencia abrumadora de que tienen una «facultad de visión» que constriñe qué clases de colores pueden ver. Ahora bien, ¿es esto así porque están siguiendo las reglas inconscientes profundas: «Si esto es infrarrojo no lo veas» o «Si esto es azul, puedes verlo»? Por lo que yo sé, jamás se ha presentado argumento alguno que muestre que las reglas de la «gramática lingüística universal» tienen un estatus diferente del de las reglas de la «gramática visual universal». Ahora bien, preguntémonos por qué exactamente no estamos tentados a decir que hay tales reglas de gramática visual universal. Después de todo, la evidencia es tan buena, de hecho es idéntica por lo que respecta a la forma, como la evidencia a favor de las reglas de la gramática lingüística universal. La respuesta, creo, es que nos es completamente obvio a partir de todo lo demás que sabemos que no hay tal nivel mental. Hay simplemente un mecanismo de *hardware* que funciona de determinadas maneras y no de otras. Estoy sugiriendo en este punto que no hay diferencia entre el estatus de la gramática visual universal inconsciente profunda y la gramática lingüística universal inconsciente profunda: ninguna de las dos existe.

Obsérvese que para rescatar el paradigma de la ciencia cognitiva no es suficiente decir que podemos decidir simplemente tratar la atribución de reglas y principios como intencionalidad *como-si*, puesto que los estados intencionales *como-si*, al no ser reales, no tienen poder causal alguno. No explican nada. El problema de la intencionalidad *como-si* no es meramente que es ubicua —que lo es—, sino que su identificación no da una explicación causal, simplemente reenuncia el problema que se supone que resuelve la atribución de intencionalidad real. Veamos cómo se aplica esto en el caso presente. Hemos intentado explicar los hechos de la adquisición del lenguaje postulando reglas de gramática universal. Si fuese verdadera, esta sería una explicación causal genuina de la adquisición del lenguaje. Pero supóngase que abandonamos esta forma de explicación y decimos simplemente que el niño actúa *como-si* estuviera siguiendo reglas, pero no lo está haciendo realmente. Si decimos esto ya no tenemos una explicación. La causa se deja ahora abierta. Hemos convertido una explicación psicológica en neurofisiología especulativa.

Si estoy en lo cierto, hemos estado cometiendo algunos errores imponentes. ¿Por qué? Creo que se debe en parte a que hemos estado suponiendo que si el *input* del sistema es significativo y el *output* lo es también, entonces todos los procesos intermedios deben serlo también. Y ciertamente hay muchos procesos significativos en la cognición. Pero allí donde no somos capaces de encontrar procesos conscientes significativos, postulamos procesos inconscientes significativos, incluso procesos inconscientes profundos. Y cuando se nos desafía invocamos el más poderoso de los argumentos: «¿Qué otra cosa podría ser?». «¿Qué otra cosa podría funcionar?» Las reglas inconscientes profundas satisfacen nuestro impulso hacia el significado y, además, ¿qué otra teoría hay? Cualquier teoría es mejor que ninguna. Una vez que cometemos esos errores nuestras teorías del inconsciente profundo ya están en marcha. Pero es simplemente falso el suponer que la significatividad del *input* y el *output* implica un conjunto de procesos significativos intermedios, y es una violación del principio de conexión el postular procesos inconscientes inaccesibles en principio.

Una de las consecuencias inesperadas de toda esta investigación es que he llegado, de manera completamente inadvertida a una defensa —si esta es la palabra correcta— del conexionismo. Entre sus otros méritos, algunos modelos conexionistas muestran al menos cómo podría un sistema convertir un *input* significativo en un *output* significativo-

tivo sin reglas, principios, inferencias u otras suertes de fenómenos significativos intermedios. Esto no equivale a decir que los modelos conexionistas existentes sean correctos —quizás todos son erróneos. Pero equivale a decir que no todos ellos son obviamente falsos o incoherentes en el modo en que lo son los modelos cognitivistas tradicionales que violan el principio de conexión.

V. Conclusión

A pesar de nuestra moderna arrogancia sobre lo mucho que sabemos, a pesar de la seguridad y universalidad de nuestra ciencia, en los asuntos que conciernen a la mente estamos característicamente confusos y en desacuerdo. Al igual que en el proverbio del ciego y el elefante, nos apoderamos de cierto presunto rasgo y lo declaramos la esencia de lo mental. «¡Ahí dentro hay oraciones invisibles!» (el lenguaje del pensamiento). «¡Ahí dentro hay un programa de ordenador!» (cognitivismo). «¡Ahí dentro hay sólo relaciones causales!» (funcionalismo). «¡Ahí dentro no hay nada!» (eliminativismo). Y así sucesivamente.

Peor aún, dejamos que nuestros métodos de investigación dicten el tema de estudio, en lugar de proceder a la inversa. Al igual que el borracho que pierde las llaves de su coche en la oscuridad de los arbustos pero las busca en la calle, a la luz de las farolas, «porque ahí hay más luz», intentamos averiguar cómo se podrían parecer los seres humanos a nuestros modelos computacionales en lugar de tratar de averiguar cómo funciona de modo efectivo la mente humana consciente. Se me pregunta frecuentemente: «¿Pero cómo puedes estudiar la conciencia *científicamente*? ¿Cómo puede haber una *teoría* sobre ella?

No creo que haya un camino simple o único hacia el redescubrimiento de la mente. Algunas indicaciones aproximadas serían:

En primer lugar, deberíamos terminar de decir cosas que son obviamente falsas. La aceptación seria de esta máxima podría revolucionar el estudio de la mente.

En segundo lugar, deberíamos recordarnos a nosotros mismos lo que sabemos con seguridad. Por ejemplo, sabemos con seguridad que dentro de nuestros cráneos hay un cerebro que algunas veces es consciente, y que los procesos cerebrales causan la conciencia en todas sus formas.

En tercer lugar, deberíamos preguntarnos a nosotros mismos qué

hechos efectivos del mundo se supone que corresponden a las afirmaciones que hacemos sobre la mente. No importa si «verdadero» significa que corresponde con los hechos, porque «corresponde con los hechos» significa corresponde con los hechos, y cualquier disciplina que aspira a describir cómo es el mundo aspira a esta correspondencia. Si uno sigue planteándose esta pregunta a la luz del conocimiento de que el cerebro es la única cosa que hay ahí dentro, y de que el cerebro causa la conciencia, creo que se topará con los resultados con los que me he topado en este libro.

Pero esto es sólo dar el primer paso en el camino de vuelta a la mente. La cuarta y última indicación sería que necesitamos redescubrir el carácter social de la mente.

BIBLIOGRAFÍA

Armstrong, D. M. (1968), A *Materialist Theory of Mind*, Routledge and Kegan Paul, Londres.

— (1980), *The Nature* of *Mind,* University of Queensland Press, Sidney.

Block, N. (1978), «Troubles with Functionalism», en *Minnesota Studies in the Philosophy of Science,* IX, University of Minnesota Press, Minneapolis, pp. 261-325.

—, ed. (1980), *Readings in Philosophy of Psychology,* vol. 1, Harvard University Press, Cambridge, Ma.

— (1990), «The Computer Model of the Mind», en D. Osherson y E. E. Smith, eds., *An Invitation to Cognitive Science,* 3, MIT Press, Cambridge, Ma., pp. 247-289.

— (inédito), «Two Concepts of Consciousness».

Block, N., y J. Fodor (1972), «What Psychological States are Not», *Philosophical Review,* 81, pp. 159-181.

Bloom, Floyd E., y Arlyne Lazerson (1988), *Brain, Mind, and Behavior,* 2.ª ed., W. H. Freeman, Nueva York.

Boolos, G. S., y R. C. Jeffrey (1989), *Computationality and Logic,* Cambridge University Press, Cambridge.

Bourdieu, P. (1977), *Outline of a Theory of Practice,* trad. ing. R. Nice, Cambridge University Press, Cambridge.

— (1990), *The Logic of Practice,* trad. ing. R. Nice, Stanford University Press, Stanford, Ca.

Changeux, J. P. (1985), *Neuronal Man: The Biology of Mind*, trad. ing. L. Garey, Pantheon Books, Nueva York.

Chisholm, R. M. (1957), *Perceiving: A Philosophical Study,* Cornell University Press, Ithaca.

Chomsky, N. (1975), *Reflections on Language,* Pantheon Books, Nueva York (hay trad. cast.: *Reflexiones sobre el lenguaje*, Ariel, Barcelona, 1979).

— (1986), *Knowledge of Language: Its Nature, Origin and Use,* Praeger Special Studies, Nueva York y Filadelfia (hay trad. cast.: *Conocimiento del lenguaje*, Alianza, Madrid, 1989).

Churchland, P. M. (1981), «Eliminative Materialism and the Propositional Attitudes», *Journal of Philosophy,* 78, pp. 67-90.
— (1984), *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind,* MIT Press., Cambridge, Ma. (hay trad. cast.: *Materia y Conciencia*, Gedisa, Barcelona, 1992).
— (1988), «The Ontological Status of Intentional States: Nailing Folk Psychology to Its Perch», *Behavioral and Brain Sciences,* 11, 3, pp. 507-508.
Churchland, P. M., y P. S. Churchland (1983), «Stalking the Wild Epistemic Engine», *Nous*, 17, pp. 5-18. Reimpreso en W. G. Lycan, ed., 1990.
Churchland, P. S. (1987), «Reply to McGinn», *Times Literary Supplement,* Cartas al director, 13 de marzo.
Davis, S., ed. (1991), *Pragmatics: A Reader*, Oxford University Press, Nueva York y Oxford.
Demopoulos, W., y R. J. Matthews (1983), «On the Hypothesis that Grammars are Mentally Represented», *Behavioral and Brain Sciences,* 6, 3, pp. 405-406.
Dennett, D. C. (1978), *Brainstorms: Philosophical Essays on Mind and Psychology*, MIT Press, Cambridge, Ma.
— (1987), *The Intentional Stance*, MIT Press, Cambridge, Ma. (hay trad. cast.: *La actitud intencional*, Gedisa, Barcelona, 1993).
— (1991), *Consciousness Explained*, Little, Brown and Company, Boston.
Dreyfus, H. L. (1972), *What Computers Can't Do*, Harper and Row, Nueva York.
— (1991), *Being-in-the-World: A Commentary On Heidegger's Being and Time Division I,* MIT Press, Cambridge, Ma.
Edelman, G. M. (1989), *The Remembered Present: A Biological Theory of Consciousness*, Basic Books, Nueva York.
Feigenbaum, E. A., y J. Feldman, eds. (1963), *Computers and Thought*, McGraw-Hill Company, Nueva York.
Feigl, H. (1958), «The "Mental" and the "Physical"», en *Minnesota Studies in the Philosophy of Science,* vol. II: *Concepts, Theories and the Mind-Body Problem*, University of Minnesota Press, Minneapolis.
Feyerabend, P. (1963), «Mental Events and the Brain», *Journal of Philosophy,* 60, pp. 295-296.
Fodor, J. (1975), *The Language of Thought*, Thomas Y. Crowell, Nueva York (hay trad. cast.: *El lenguaje del pensamiento*, Alianza, Madrid, 1984).
— (1986), «Banish DisContent», en Butterfield, J., ed. *Language Mind and Logic*, Cambridge University Press, Cambridge.
— (1987), *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press, Cambridge, Ma. (hay trad. cast.: *Psicosemántica*, Tecnos, Madrid, 1994).
Foucault, M. (1972), *The Archaeology of Knowledge*, trad. ing. A. M. Sheridan Smith, Harper and Row, Nueva York.

Freud, S. (1895), «Project for Scientific Psychology», en *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, vol. 1, trad. ing. James Strachey, Hogarth Press, Londres, 1966, pp. 295-343.
— (1915), «The Unconscious in Psychoanalysis», en *Collected Papers,* vol. 4, trad. ing. J. Riviere, Basic Books, Nueva York, 1959, pp. 98-136.
— (1949), *Outline of Psychoanalysis.* trad. ing. James Strachey, Hogarth Press, Londres.
Gardner, H. (1985), *The Mind's New Science: A History of the Cognitive Revolution*, Basic Books, Nueva York (hay trad. cast.: *La nueva ciencia de la mente*, Paidós, Barcelona, 1988).
Gazzaniga, M. S. (1970), *The Bisected Brain*, Appleton Century Crofts, Nueva York.
Geach, P. (1957), *Mental Acts*, Routledge y Kegan Paul, Londres.
Grice, P. (1975), «Method in Philosophical Psychology (From the Banal to the Bizarre)», *Proceedings and Addresses of the American Philosophical Association*, vol. 48, noviembre de 1975, pp. 23-53.
Griffin, D. R (1981), *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience*, Rockefeller University Press, Nueva York.
Hampshire, S. (1950), «Critical Notice of Ryle, *The Concept of Mind*», *Mind*, LIX, 234, pp. 237-255.
Hare, R M. (1952), *The Language of Morals*, Oxford University Press, Oxford.
Haugeland, J., ed. (1981), *Mind Design*, MIT Press, Cambridge, Ma.
— (1982), «Weak Supervenience», *American Philosophical Quarterly,* 19, 1, pp. 93-104.
Hempel, C. G. (1949), «The Logical Analysis of Psychology», en H. Feigl y W. Sellars, eds., *Readings in Philosophical Analysis*, Appleton Century Crofts, Nueva York.
Hobbs, J. R. (1990), «Matter, Levels, and Consciousness», *Behavioral and Brain Sciences,* 13, 4, pp. 610-611.
Horgan, T., y J. Woodward (1985), «Folk Psychology is Here to Stay», *Philosophical Review,* XCIV, 2, pp. 197-220.
Jackendoff, R. (1987), *Consciousness and the Computational Mind*, MIT Press, Cambridge, Ma.
Jackson, F. (1982), «Epiphenomenal Qualia», *Philosophical Quarterly,* 32, pp. 127-136.
Johnson-Laird, P. N. (1983), *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, Harvard University Press, Cambridge, Ma.
— (1988), *The Computer and the Mind*, Harvard University Press, Cambridge, Ma. (hay trad. cast.: *El ordenador y la mente*, Paidós, Barcelona, 1990).
Kim, J. (1979), «Causality, Identity and Supervenience in the Mind-Body Problem», *Midwest Studies in Philosophy,* 4, pp. 31-49.

— (1982), «Psychophysical Supervenience», *Philosophical Studies,* 41, 1, pp. 51-70.

Kripke, S. A. (1971), «Naming and Necessity», en D. Davidson y G. Harman, eds., *Semantics of Natural Language*, Reidel, Dordrecht, pp. 253-355 y 763-769.

— (1982), *Wittgenstein on Rules and Private Language*, Basil Blackwell, Oxford (hay trad. cast.: *Wittgenstein sobre reglas y lenguaje privado*, UNAM, Instituto de Investigaciones Filosóficas, México, 1992).

Kuffler, S. W., y J. G. Nicholls (1976), *From Neuron to Brain*, Sinauer Associates, Sunderland, Ma.

Lashley, K. (1956), «Cerebral Organization and Behavior», en H. Solomon, S. Cobb y W. Penfield, eds., *The Brain and Human Behavior*, Williams and Wilkins Press, Baltimore.

Lepore, E., y R. van Gulick, eds. (1991), *John Searle and His Critics*, Basil Blackwell, Oxford.

Lettvin, J. Y., H. R. Maturana, W. S. McCuloch y W. H. Pitts (1959), «What the Frog's Eye Tells to the Frog's Brain», *Proceedings of the Institute of Radio Engineers,* 47, 1940-1951. Reimpreso en W. S. McCulloch (1965).

Lewis, D. (1966), «An Argument for the Identity Theory», *Journal of Philosophy,* 63,1, pp. 17-25. Reimpreso en D. Rosenthal, ed., 1971.

— (1972), «Psychological and Theoretical Identification», *Australasian Journal of Philosophy,* 50, pp. 249-258.

Lisberger, S. G. (1988), «The Neural Basis for Learning of Simple Motor Skills», *Science,* 4, 242, pp. 728-735.

—, y T. A. Pavelco (1988), «Brain Stem Neurons in Modified Pathways for Motor Learning in the Primate Vestibulo-Ocular Reflex», *Science*, 4, 242, pp. 771-773.

Lycan, W. G. (1971), «Kripke and Materialism», *Journal of Philosophy,* 71, 18, pp. 677-689.

— (1987a), *Consciousness,* MIT Press, Cambridge, Ma.

— (1987b), «What is the "Subjectivity" of the Mental», *Philosophical Perspectives,* 4, pp. 109-130.

—, ed. (1990), *Mind and Cognition: A Reader,* Basil Blackwell, Cambridge, Ma.

Marr, D. (1982), *Vision*, W. H. Freeman and Company, San Francisco.

McCulloch, W. S. (1965), *The Embodiment of Mind,* Harvard University Press, Cambridge, Ma.

McGinn, C. (1977), «Anomalous Monism and Kripke's Cartesian Intuitions», *Analysis,* 37, 2, pp. 78-80.

— (1987), «Review of P. S. Churchland, *Neurophilosophy*», *Times Literary Supplement,* 6 de febrero, pp. 131-132.

— (1991), *The Problem of Consciousness,* Basil Blackwell, Oxford.

Millikan, R. (1984), *Language, Thought and Other Biological Categories: New Foundations for Realism*, MIT Press, Cambridge, Ma.

Minsky, M. L. (1986), *Society of Mind,* Simon and Schuster, Nueva York (hay trad. cast.: *La sociedad de la mente*, Galápago, Buenos Aires, 1986).

Moore, G. E. (1922), *Philosophical Studies,* Routledge and Kegan Paul, Londres.

Nagel, T. (1974), «What Is It Like to Be a Bat?», *Philosophical Review,* 4, LXXXIII, pp. 435-450.

— (1986), *The View from Nowhere,* Oxford University Press, Oxford.

Newell, A. (1982), «The Knowledge Level», *Artificial Intelligence,* 18, pp. 87-127.

Ogden, C. K., e I. A. Richards (1926), *The Meaning of Meaning,* Harcourt, Brace and Company, Londres (hay trad. cast.: *El significado de significado,* Paidós, Buenos Aires, 1954).

Penfield, W. (1975), *The Mystery of the Mind: A Critical Study of Consciousness and the Human Brain,* Princeton University Press, Princeton.

Penrose, R. (1989), *The Emperor's New Mind,* Oxford University Press, Oxford (hay trad. cast.: *La nueva mente del emperador*, Mondadori, Madrid, 1991).

Place, U. T. (1956), «Is Consciousness a Brain Process?», *British Journal of Psychology,* 47, pp. 44-50.

— (1988), «Thirty Years On—Is Consciousness Still a Brain Process?», *Australasian Journal of Philosophy*, 66, 2, pp. 208-219.

Postman, L., J. Bruner y R. Walk (1951), «The Perception of the Error», *British Journal of Psychology,* 42, pp. 1-10.

Putnam, H. (1960), «Minds and Machines», en S. Hook, ed., *Dimensions of Mind*, Collier Books, Nueva York.

— (1963), «Brains and Behavior», en R. Butler, ed., *Analytical Philosophy,* Basil Blackwell, Oxford.

— (1967), «The Mental Life of Some Machines», en H. Castañeda, ed., *Intentionality, Minds, and Perception,* Wayne State University Press, Detroit, Mi.

— (1975a), «Philosophy and Our Mental Life», en *Mind, Language and Reality: Philosophical Papers*, vol. 2, Cambridge University Press, Cambridge.

— (1975b), «The Meaning of "Meaning"», en K. Gunderson, ed., *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science,* VII, University of Minnesota Press, Minneapolis (hay trad. cast.: «El significado de *significado*», Valdés, 1991, pp.131-194).

Pylyshyn, Z. W. (1984), *Computation and Cognition: Toward a Foundation for Cognitive Science,* MIT Press, Cambridge, Ma.

Quine, W. V. O. (1960), *Word and Object,* MIT Press, Cambridge, Ma. (hay trad. cast. de Manuel Sacristán: *Palabra y objeto*, Labor, Barcelona, 1968).

Rey, G. (1983), «A Reason for Doubting the Existence of Consciousness», en R. Davidson, G. Schwartz, D. Shapiro, eds., *Consciousness and Self-Regulation,* 3, Plenum, Nueva York, pp. 1-39.

— (1988), «A Question about Consciousness», en H. Otto, J. Tuedio, eds., *Perspectives on Mind,* Reidel, Dordrecht.

Rock, I. (1984), *Perception,* Scientific American Library, W. H. Freeman, Nueva York.

Rorty, R. (1965), «Mind-Body Identity, Privacy and Categories», *Review of Metaphysics,* 29, 1, pp. 24-54.

— (1970), «Incorrigibility as the Mark of the Mental», *Journal of Philosophy,* LXVII, 12, pp. 399-424.

— (1979), *Philosophy and the Mirror of Nature,* Princeton University Press, Princeton (hay trad. cast.: *Filosofía y el espejo de la naturaleza,* Cátedra, Madrid, 1983).

Rosenthal, D., ed. (1971), *Materialism and the Mind-Body Problem,* Prentice Hall, Englewood Cliffs, N.J.

— (1991), *The Nature of Mind,* Oxford University Press, Nueva York.

Ryle, G. (1949), *The Concept of Mind,* Barnes and Noble, Nueva York (hay trad. cast.: *El concepto de lo mental,* Paidós, Buenos Aires, 1967).

Sacks, O. (1985), *The Man Who Mistook His Wife For a Hat: And Other Clinical Tales,* Simon and Schuster, Nueva York.

Sarna, S.K., y M. F. Otterson (1988), «Gastrointestinal Motility: Some Basic Concepts», *Pharmacology: Supplement,* 36, pp. 7-14.

Schiffer, S. R. (1987), *Remnants of Meaning,* MIT Press, Cambridge, Ma.

Searle, J. R. (1976), «The Rules of the Language Game», recensión de Noam Chomsky, *Reflections on Language, The Times Literary Supplement,* 10 de septiembre.

— (1978), «Literal Meaning», *Erkenntnis,* 1, pp. 207-224. Reimpreso en Searle (1979).

— (1979), *Expression and Meaning,* Cambridge University Press, Cambridge.

— (1980a), «Minds, Brains, and Programs», *Behavorial and Brain Sciences,* 3, pp. 417-424.

— (1980b), «Intrinsic Intentionality: Reply to Criticisms of Minds, Brains and Programs», *Behavioral and Brain Sciences,* 3, pp. 450-456.

— (1980c), «The Background of Meaning», en J. R Searle, F. Kiefer y M. Bierwisch, eds., *Speech Act Theory and Pragmatics,* Reidel, Dordrecht,

— (1982), «The Chinese Room Revisited: Response to Further Commentaries on "Minds, Brains, and Programs"», *Behavioral and Brain Sciences,* 5, 2, pp. 345-348.

— (1983), *Intentionality: An Essay in the Philosophy of Mind,* Cambridge University Press, Cambridge (hay trad. cast.: *Intencionalidad,* Tecnos, Madrid, 1992).

— (1984a), «Intentionality and Its Place in Nature», *Synthese,* 61, pp. 3-16.

— (1984b), *Minds, Brains, and Science: The 1984 Reith Lectures,* Harvard University Press. Cambridge, Ma. (hay trad. cast.: *Mentes, cerebros y ciencia,* Cátedra, Madrid, 1986).

— (1987), «Indeterminacy, Empiricism, and the First Person», *Journal of Philosophy,* LXXXIV, 3, pp. 123-146.
— (1990), «Collective Intentionality and Action», en P. Cohen, J. Morgan y M. E. Pollack, eds., *Intentions in Communications*, MIT Press, Cambridge, Ma.
— (1991), «Response: The Background of Intentionality and Action», en E. Lepore y R. van Gulick, eds. (1991), pp. 289-299.
— (inédito), «Skepticism about Rules and Intentionality».
Segal, G. (1991), «Review of Garfield, J., *Belief in Psychology*», *Philosophical Review* C, 3, pp. 463-466.
Shaffer, J. (1961), «Could Mental States be Brain Processes?», *Journal of Philosophy,* 58, 26, pp. 813-822.
Sharples, M., D. Hogg, C. Hutchinson, S. Torrence y D. Young (1988), *Computers and Thought: A Practical Introduction to Artificial Intelligence,* MIT Press, Cambridge, Ma.
Shepherd, G. M. (1983), *Neurobiology*, Oxford University Press, Oxford y Nueva York.
Sher, G. (1977), «Kripke, Cartesian Intuitions, and Materialism», *Canadian Journal of Philosophy,* 7.
Smart, J. J. C. (1959), «Sensations and Brain Processes», *Philosophical Review,* 68, pp. 141-156.
Smith, D. W. (1986), «The Structure of (Self-)Consciousness», *Topoi,* 5, 2, pp. 149-156.
Sober, E. (1984), *The Nature of Selection: Evolutionary Theory in Philosophical Focus*, MIT Press, Cambridge, Ma.
Stevenson, J. T. (1960), «Sensations and Brain Processes: A Reply to J. J. C. Smart», *Philosophical Review,* 69, pp. 505-510.
Stich, S. P. (1983), *From Folk Psychology to Cognitive Science: The Case against Belief*, MIT Press, Cambridge, Ma.
— (1987), «Review of Searle, J., *Minds, Brains and Science*», *Philosophical Review,* XCVI, 1, pp. 129-133.
Turing, A. (1950), «Computing Machinery and Intelligence», *Mind,* 59, pp. 433-460 (hay trad. cast.: «¿Puede pensar una máquina?», en *Mentes y máquinas*, Tecnos, Madrid, 1985).
Valdés Villanueva, Luis M. (1991), *La búsqueda del significado*, Tecnos, Madrid.
Waldrop, M. M. (1988), «Toward a Unified Theory of Cognition», y «SOAR: A Unified Theory of Cognition», *Science,* 241 (1 de julio), pp. 27-29, y (15 de julio) pp. 296-298.
Watson, J. B. (1925), *Behaviorism,* Norton Press, Nueva York (hay trad. cast. en Paidós, Buenos Aires, 1972).
Weiskrantz, L. *et al.* (1974), «Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation», *Brain,* 97, pp. 709-728.
Williams, B. (1987), «Leviathan's Program: Review of Marvin Minsky, *The Society of Mind*», *The New York Review of Books*, 11 de junio.

Wittgenstein, L. (1953), *Philosophical Investigations*, Basil Blackwell, Oxford (hay trad. cast.: *Investigaciones filosóficas*, UNAM/Crítica, Barcelona, 1988).

— (1969), *On Certainty*, Basil Blackwell, Oxford (hay trad. cast.: *Sobre la certeza*, Gedisa, Barcelona, 1988).

ÍNDÍCE ANALÍTICO

ÍNDÍCE ONOMÁSTICO



ÍNDICE